HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

BACHELORARBEIT

Herr
Silvio Oswald

Energieprofilbasierende Analysen, Clustering und Detektion struktureller bzw. funktioneller Proteinmotive und Faltungsklassen

Mittweida, 2012

Fakultät Mathematik, Naturwissenschaften, Informatik

BACHELORARBEIT

Energieprofilbasierende Analysen, Clustering und Detektion struktureller bzw. funktioneller Proteinmotive und Faltungsklassen

Autor:
Herr

Silvio Oswald

Studiengang:
Biotechnologie/Bioinformatik

Seminargruppe:
BI09w2-B

Erstprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer:
B.Sc. Florian Heinke

Einreichung:
Mittweida, 22.8.2012

Verteidigung/Bewertung:
Mittweida, 2012

Faculty Mathematics, Natural Sciences, Informatics

BACHELORTHESIS

Energyprofile-based Analyses, Clustering and Detection of structural and functional Proteinmotifs and Foldclasses

author:
Mr.

Silvio Oswald

course of studies:
Biotechnology/Bioinformatics

seminar group:
BI09w2-B

first examiner:
Prof.Dr. rer. nat. Dirk Labudde

second examiner:
B.Sc. Florian Heinke

submission:
Mittweida, August the 22th, 2012

defence/ evaluation:

Mittweida, 2012

Bibliografische Beschreibung:

Oswald, Silvio:
Energieprofilbasierende Analysen, Clustering und Detektion struktureller bzw. funktioneller Proteinmotive und Faltungsklassen. - 2012. – 10, 79, 8 S.
Mittweida, Hochschule Mittweida, Fakultät Fakultät Mathematik/Naturwissenschaften/Informatik, Bachelorarbeit, 2012

Referat:

In der vorliegenden Arbeit werden strukturelle und funktionelle Proteinmotive hinsichtlich ihrer energetischen Charakteristika untersucht und nach energetischen Abständen über hierarchische Clusterverfahren geclustert. Dabei sollen Gesetzmäßigkeiten offengelegt werden, die sich über die Abstraktionsebenen der Sequenz, Struktur, Funktion und der Energie erstrecken.

Danksagung

Da diese Arbeit nicht allein mein Schaffen ist, möchte ich an erster Stelle einer Reihe von Menschen danken. An erster Stelle danke ich Prof. Dr. Dirk Labudde, der es mir ermöglichte an diesem für mich sehr interessanten Thema, eingebettet in eine Arbeitsgruppe, zu arbeiten. Er half uns mit seiner Betreuung über die Zeit der Projektarbeit, des Praxismoduls und zu guter Letzt der Bachelorphase unsere Arbeit in einem größeren Kontext zu verstehen.
Ein nicht minder großer Dank gilt meinem Betreuer B.Sc. Florian Heinke, der zu jeder Zeit mit Rat und Tat bereit war sich kleinen und großen Problemen anzunehmen und mir oft Ideen vermittelt hat, diese zu lösen. Von diesen Dingen werde ich eine Menge auf meinem weiteren Weg mitnehmen.
Weiterhin danke ich meinen Kommilitonen Alexander Hampel und Mathias Langer für oft anstrengende, aber zumeist fruchtbare Diskussionen über verschiedenste Thematiken unseres Fachbereiches. Auf diesem Weg sind möglicherweise mehr Ideen und Einsichten zustande gekommen, als einem selbst bewusst ist.
Zu guter Letzt danke ich auch meinem Vater Volkmar Oswald, der sich die Zeit und vor allem die Nerven nahm diese Arbeit auf orthographische und grammatikalische Fehltritte hin zu überprüfen.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CATH	Class, Architecture, Topology, Homology
PDB	Protein Data Bank
SCOP	Structural Classification of Proteins
DNA	Desoxyribonucleinsäure
ATP	Adenosintriphosphat
GTP	Guanosintriphosphat

1 Einleitung

Auch nach einem halben Jahrhundert Forschungs- und Denkarbeit auf dem weiten Feld der Molekularbiologie ist es eines der größten ungelösten Probleme, wie die Beziehung zwischen Sequenz, 3D-Struktur und natürlicher Funktion von Proteinen in Wirklichkeit aussieht.

Das größte Problem stellt dabei einerseits die Erklärung von Faltungsprozessen von Proteinen, andererseits das Wesen ihrer katalytischen Wirkung dar. Ein Ansatz, der es an dieser Stelle ermöglicht die Natur von Proteinstrukturen intensiver zu beschreiben, als dies Sequenzen und aus ihnen gewonnene Metadaten können, sind sogenannte Energieprofile [1]. Dabei wird die Wechselwirkung jeder Aminosäure im Protein zu anderen Aminosäuren in der Struktur berücksichtigt. Es wird dabei für jede Aminosäure eine freie Energie berechnet, die sich relativ zur Position dieser in der Struktur verhält. Weiterhin fließen individuelle physiko-chemische und konformationelle Eigenschaften jeder Aminosäure in die Berechnung des Energieprofils ein.

Proteindatenbanken wie die PDBeMotif oder die Prosite beinhalten Daten über das Auftreten von strukturellen bzw. funktionellen Proteinmotiven. Strukturelle Motive sind dabei als Strukturfragmente eines Proteins zu verstehen, die sich durch besondere strukturelle Konformationseigenschaften auszeichnen [2]. Hierbei spielen Winkelkonformationen und Wasserstoffbrückenbindungen eine Rolle.

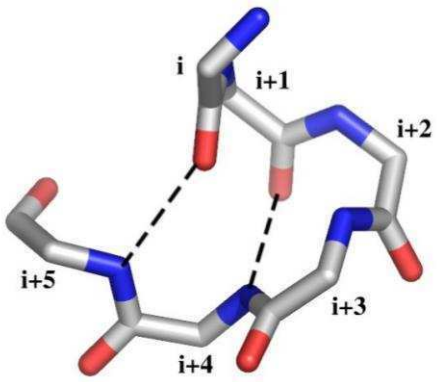


Abbildung 1: Schellmannloop [2]

Der Schellmannloop stellt einen Vertreter der Strukturmotive dar. Mit einer Länge von sechs Aminosäuren ist er eines der längsten betrachteten Strukturmotive.

Getrennt von den strukturellen Motiven sind funktionelle Motive zu betrachten, die sich durch eine eindeutig definierte Funktion auszeichnen. Beispiele hierfür sind ligandenspezifische Glykolysierungs- oder Phosphorylierungsstellen.

Abbildung 2: N-Glykolysierungsstelle in 1B9W

Glykolysierungsstellen (rot hervorgehoben) kommen in globulären Proteinen häufig vor. Sie sind dafür zuständig Kohlenhydratketten an Proteine oder Lipide zu binden. Dieser Reaktionsweg stellt eine der wichtigsten posttranslationalen Modifikationen für Proteine dar.

Ein Ziel dieser Arbeit ist es strukturelle beziehungsweise funktionelle Motive auf Ebene der Energieprofile zu untersuchen. Das Hauptaugenmerk liegt dabei darauf eine Klassifizierbarkeit dieser Motive aus energetischen Charakteristika abzuleiten. Dabei soll auch die Frage diskutiert werden, ob es möglich ist die betrachteten Motive allein aus Energieprofilinformationen effizient zu detektieren.

Grundsätzlich können Proteine in globuläre, also frei im Cytoplasma bzw. dem entsprechenden Zellorganell befindlichen Proteinen und Membranproteinen unterschieden werden. In der vorliegenden Arbeiten sollen nur globuläre Proteine untersucht werden.

2 Theoretische Grundlagen

Für das Verständnis der später folgenden Ausführungen sollen in diesem Kapitel einige Grundlagen über Proteinstrukturen, Proteinfaltung, Proteinmotive und Energieprofile gelegt werden.

2.1 Proteine

Der als Vater der modernen Chemie geltende schwedische Chemiker Jöns Jakob Berzelius schlug im Jahr 1838 basierend auf dem griechischen πρωτεῖος proteios für ‚grundlegend' oder ‚vorrangig' erstmals das Wort Protein vor.

Hinter Berzelius' Vorschlag verbarg sich die, heute als falsch erwiesene These, dass alle Proteine auf einer gemeinsamen Grundsubstanz basieren. Berzelius hatte insofern damit Unrecht, dass zwar alle Proteine auf einer gemeinsamen Stoffklasse basieren, jedoch nicht auf ein und demselben Stoff. Seit 1950 – dem Jahr, in dem Pehr Edman seine als Edman-Abbau bekannte Proteinsequenziertechnik vorstellte – ist bekannt, dass Proteine linear aufgebaute Makromoleküle sind, die sich aus 20 voneinander unterscheidbaren kanonischen Aminosäuren zusammensetzen.

Als wesentlicher Bestandteil jeder der uns bekannten Zellen fungieren sie als Stoffwechselbestandteile, Stabilisierungselemente und Signalüberträger. Sie sind beispielsweise dafür verantwortlich, die in Form von DNA abgelegte Erbinformation zu replizieren, sowie eventuelle Fehler zu korrigieren. Beinahe jede zelluläre Funktion wird erst dadurch ermöglicht, dass ein Protein – in diesem Fall speziell die Enzyme – eine chemische Reaktion katalysieren. Andere Proteine, wie das Kinesin wirken als molekulare Motoren und bewegen Zellorganellen durch das Cytoplasma [4].

Die immense Fülle an Funktionen, die Proteine erfüllen ist ein Resultat der Differenzierung ihrer Struktur. Um also zu verstehen, warum Proteine so vielseitig sind, ist es vonnöten, sich mit dem Wesen ihrer Struktur auseinanderzusetzen.

2.2 Aufbau von Proteinstrukturen

2.2.1 Primärstruktur

Die sogenannte Primärstruktur beschreibt die erste Abstraktionsebene von Proteinen und beschreibt die lineare Abfolge der Aminosäuren, die das Protein bilden. Aminosäuren sind eine Klasse organischer Kohlenstoffverbindungen, die sich aus mindestens einer Carboxyl- (COOH), einer Aminogruppe (NH_2) und einem individuellen organischen Rest zusammensetzen.

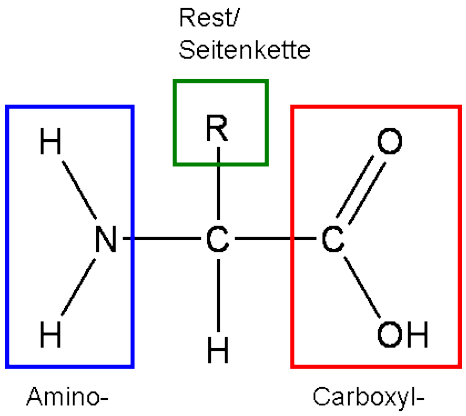


Abbildung 3: Aufbau von Aminosäuren [5]

Jede Aminosäure orientiert sich an dieser Grundstruktur, wobei R ein aminosäurenspezifischer organischer Rest ist.

Das an am Sauerstoffatom gebundene Wasserstoffatom der Carboxylgruppe wird durch die hohe Elektronegativität des Sauerstoffatoms partiell positiv geladen, was Reaktionen an dieser Stelle ermöglicht. Im Gegensatz dazu ist die Aminogruppe in der Lage andere partiell positiv geladene Wasserstoffatome aufzunehmen. Dieser Fakt ist, wie sich später zeigen wird essentiell für die Bildung größerer Aminosäureketten.

Der individuelle Rest einer Aminosäure bestimmt ihre jeweiligen chemischen Eigenschaften wie Ladung, Hydrophobizität oder durch seine Größe die Größe der gesamten Aminosäure. Ist beispielsweise der Rest einer Aminosäure ein hydrophobes (wasserabweisendes) Teilmolekül, so verhält sich die gesamte Aminosäure hydrophob.

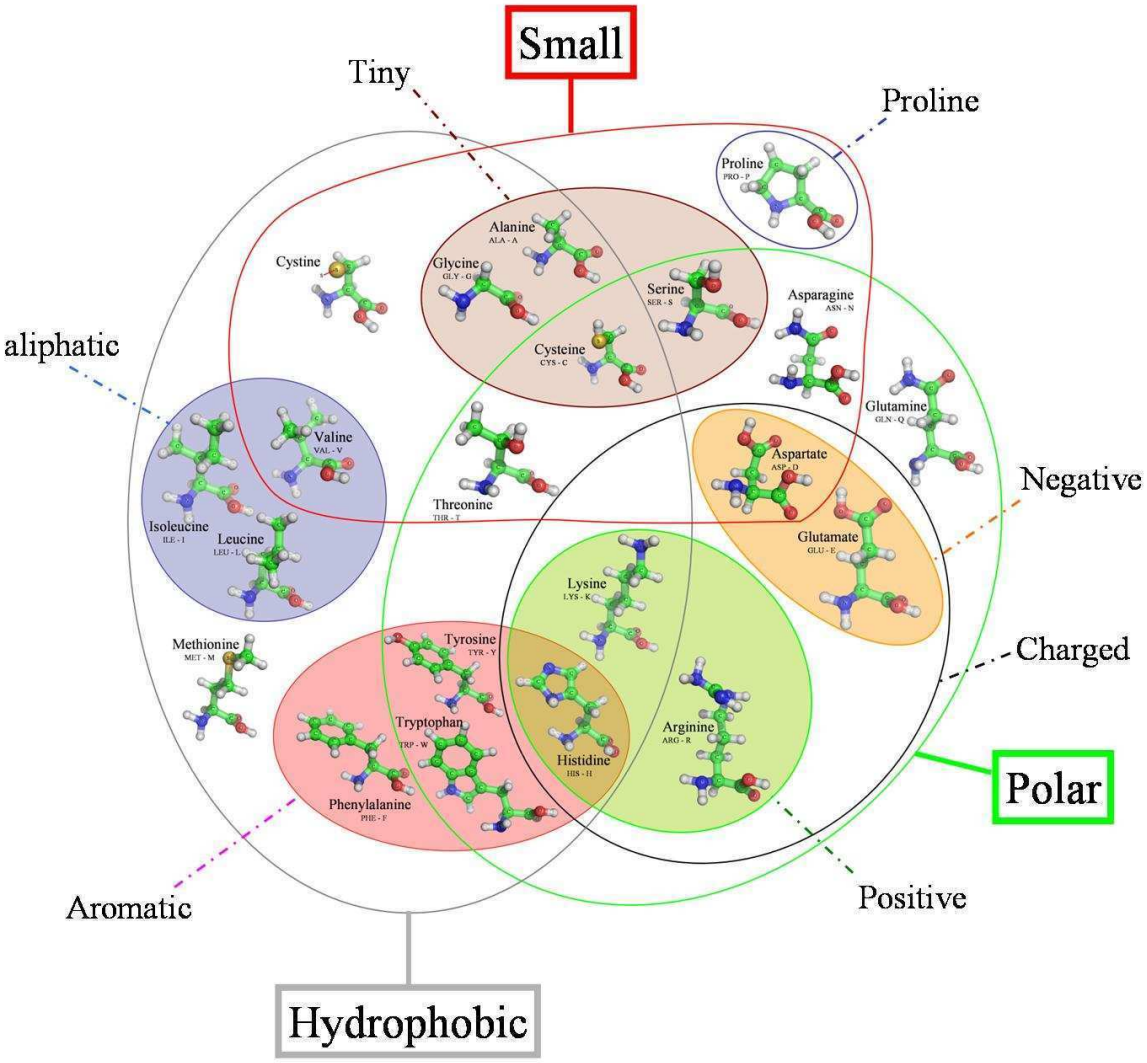


Abbildung 4: Venn-Diagramm der Aminosäuren [6]

Aufgrund der chemischen Unterschiede der Restgruppe jeder Aminosäure lassen sich die 20 proteinogenen Aminosäuren in diverse Gruppen einordnen.

In der obigen Abbildung sind die 20 kanonischen Aminosäuren anhand ihrer Atome, aus denen sie aufgebaut sind, eingefärbt. Eine grüne Farbe stellt dabei ein Kohlenstoff-, eine weiße ein Wasserstoff-, eine blaue ein Stickstoff-, eine rote ein Sauerstoff- und eine gelbe ein Schwefelatom dar.

Eigenschaften wie die Hydrophobizität, also der Charakter einer Aminosäure wasserlöslich oder eher wasserunlöslich zu sein, spielen einerseits bei der Bildung komplexerer Proteinstrukturen und andererseits bei der Ausübung der Funktion des Proteins eine elementare Rolle.

Die Primärstruktur, die die einfachste Stufe der Strukturbildung eines Proteins darstellt, wird durch eine Kondensationsreaktion mehrerer Aminosäuren und die daraus resultierende Bildung eines Polypeptids determiniert.

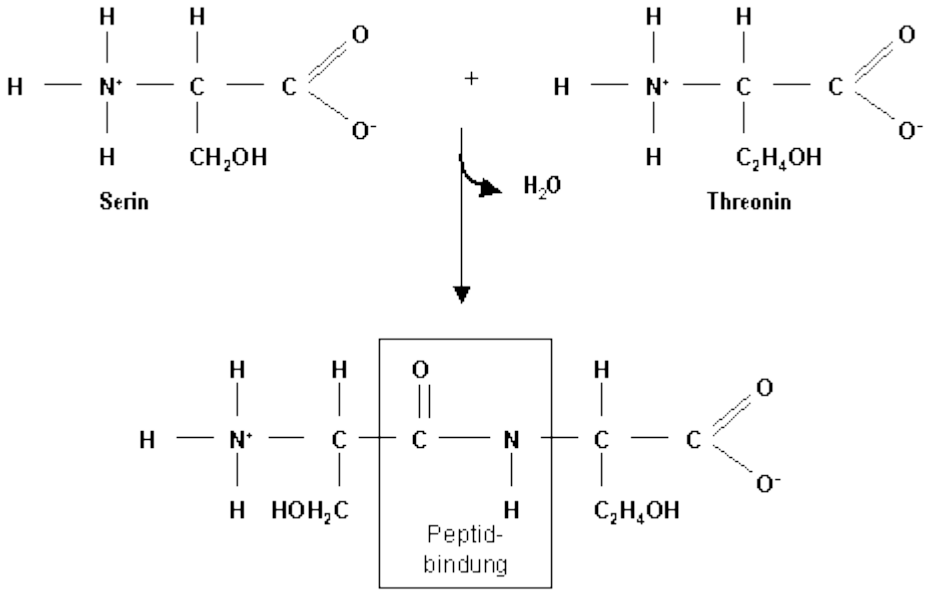


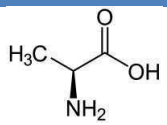
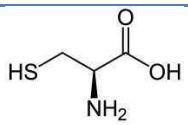
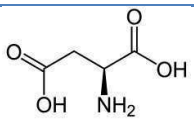
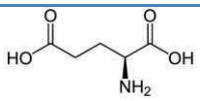
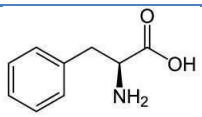
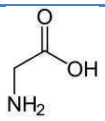
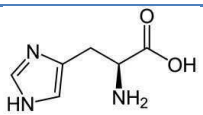
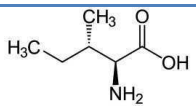
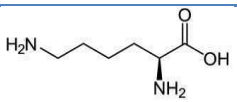
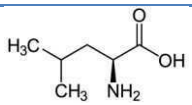
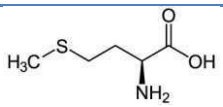
Abbildung 5: Bildung einer Peptidbindung [7]

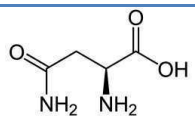
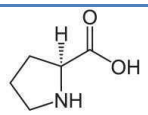
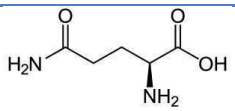
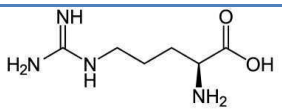
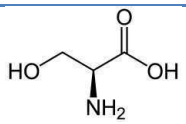
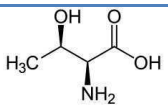
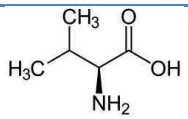
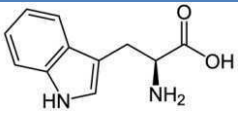
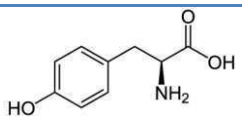
Die Peptidbindung entsteht zwischen der Carboxyl- und der Aminogruppe zweier Aminosäuren

Hier dargestellt ist eine einstufige Kondensationsreaktion für zwei Aminosäuren, die nach Reaktionsabschluss eine gemeinsame Peptidbindung besitzen. Auf diese Weise können prinziell beliebig viele Aminosäuren zu einem Polypeptid verbunden werden. Bei der Bildung dieser Bindung wird ein Wassermolekül frei, um sie wieder zu lösen würde wieder ein Wassermolekül benötigt werden, um die Aminosäuren wieder in ihren stabilen Einzelzustand zurückzuführen.

Die Primärstruktur stellt sich folglich als Abfolge von Aminosäuren, die als Sequenz des Proteins bezeichnet wird, dar. Für die Beschreibung einer Sequenz wurde ein einheitlicher Formalismus – der Ein- bzw. Drei-Letter-Code – eingeführt, der es erlaubt die Primärstruktur eines Proteins auf eine Zeichenkette zu reduzieren.

Tabelle 1: Die 20 kanonischen Aminosäuren

Aminosäure	Ein-Letter-Code	Drei-Letter-Code	Strukturformel	Relative Häufigkeit in Proteinen [8]
Alanin	A	Ala		9,0 %
Cystein	C	Cys		2,8 %
Asparaginsäure	D	Asp		5,5 %
Glutaminsäure	E	Glu		6,2 %
Phenylalanin	F	Phe		3,5 %
Glycin	G	Gly		7,5 %
Histidin	H	His		2,1 %
Isoleucin	I	Ile		4,6 %
Lysin	K	Lys		7,0 %
Leucin	L	Leu		7,5 %
Methionin	M	Met		1,7 %

Asparagin	N	Asn		4,4 %
Prolin	P	Pro		4,6 %
Glutamin	Q	Gln		3,9 %
Arginin	R	Arg		4,7 %
Serin	S	Ser		7,1 %
Threonin	T	Thr		6,0 %
Valin	V	Val		6,9 %
Tryptophan	W	Trp		1,1 %
Tyrosin	Y	Tyr		3,5 %

2.2.2 Sekundärstruktur

Eine Besonderheit der Peptidbindung ist, dass es sich bei ihr nicht um eine starre Bindung, sondern vielmehr um ein drehbares Konstrukt handelt.

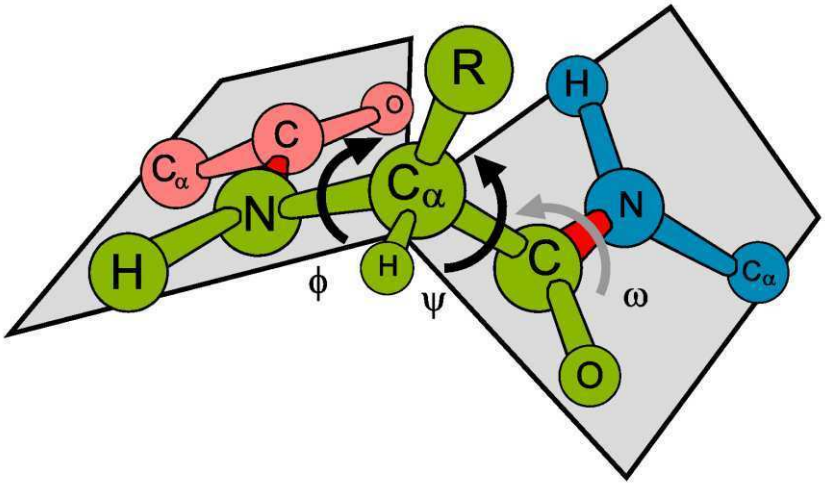


Abbildung 6: Torsionswinkel innerhalb der Peptidbindung [9]

Die Ausrichtung der Aminosäuren innerhalb des Polypeptids ist durch drei sogenannte Torsionswinkel bestimmt. Sie ermöglichen eine Drehung der Bindungen um das zentrale C_α-Atom herum.

Das Auftreten verschiedener Winkelkonformationen innerhalb einer Primärstruktur ist ein weiterer Grund dafür, dass sich Proteinstrukturen mannigfaltig differenzieren und sich verschiedene Sekundärstrukturelemente beschreiben lassen.

Diese Sekundärstrukturelemente stellen die nächsthöhere Abstraktionsebene für Proteinstrukturen – die Sekundärstruktur – dar. Es handelt sich dabei um regelmäßige Substrukturen, die sich dreidimensional entlang des Hauptkettenverlaufs (backbone) der Primärstruktur bilden [10]. Ein Vergleich der dreidimensionalen Struktur vieler sequenziell unterschiedlicher Proteine zeigt, dass zwei Sekundärstrukturelemente häufig in Teilbereichen vorkommen, obwohl die Gesamtkonformation jedes Proteins einzigartig ist: α-Helices (siehe Abbildung 8) und β-Sheets (siehe Abbildung 9) [4]. Bereiche in der Proteinstruktur, die keinem dieser beiden Strukturelemente zuzuordnen sind, werden als Random Coils bezeichnet. Welches von beiden Sekundärstrukturelementen in einem Bereich der backbone vorliegt, wird durch die intermolekularen Kräfte zwischen Aminosäuren und der daraus folgenden räumlichen Anordnung dieser determiniert [10].

Sind beispielsweise die φ- und ψ-Torsionswinkel einer Teilkette im Protein über mehr als 3,6 Aminosäuren konstant, so bildet sich eine α-Helix aus [11]. Innerhalb der Helix werden die Aminosäuren durch zusätzliche Wasserstoffbrücken stabilisiert.

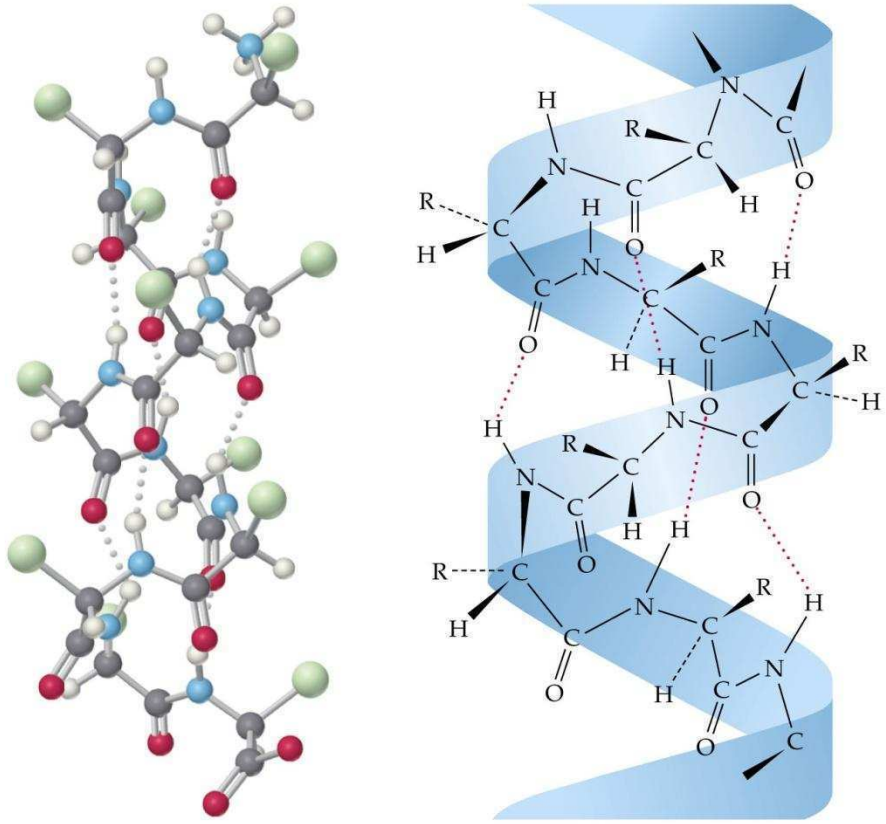


Abbildung 7: Aufbau einer α-Helix [12]

Zylindrische Strukturen wie eine α-Helix bilden sich aus, wenn sich eine Polypeptidkette mehrfach um die eigene Achse dreht. Zwischen jeder vierten Peptidbindung wird eine Wasserstoffbrücke (gepunktet dargestellt) ausgebildet [4].

Helices treten besonders häufig in Proteinen auf, die in Zellmembranen lokalisiert sind. An diesen Stellen durchqueren sie als sogenannte Transmembranhelices die Lipiddoppelschicht der Biomembran. Ein Sonderfall unter den Helixstrukturen stellt die coiled-coil-Struktur dar, bei der sich zwei Helices umeinander winden und so besonders stabile Strukturen ausbilden. Die Aufgabe solcher Strukturen zeigt sich beispielsweise beim Myosin, dass ein zentraler Bestandteil des Muskelkontraktionsprozesses ist [4].

Anders als im Falle der α-Helix, bei der eine bestimmte Anzahl an Aminosäuren einen konstanten Torsionswinkel aufweisen muss, damit die Struktur entsteht verhält es sich bei β -strands. In einem β-strand weisen die Torsionswinkel der Aminosäuren innerhalb der Polypeptidkette ein alternierendes Verhalten auf. Diese Eigenschaft zeigt sich in der Struktur als ein nahezu ebenes Gebilde, wobei mehrere parallel oder antiparallel verlaufende β-strands ein β-sheet oder β-Faltblatt bilden können.

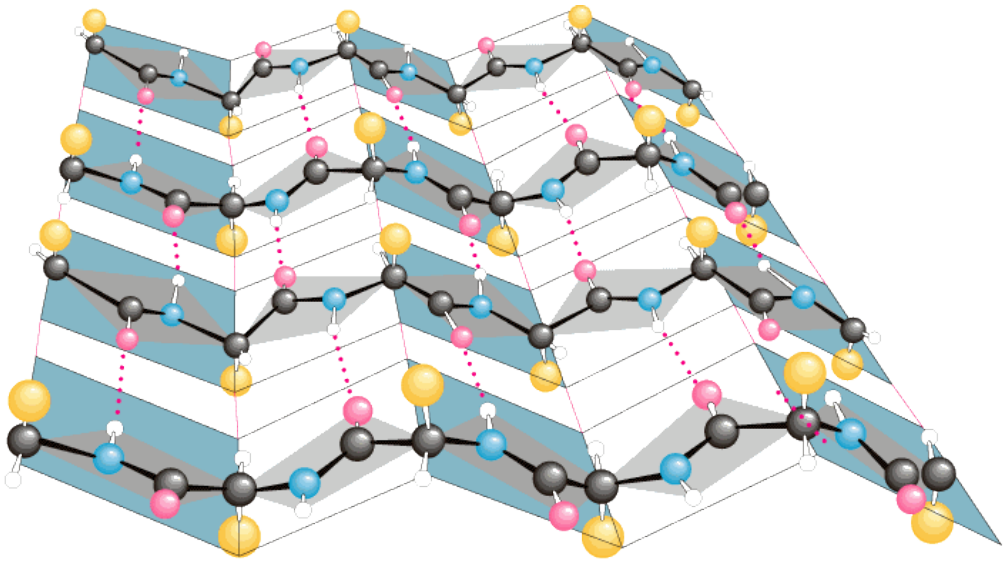


Abbildung 8: Struktur eines β-Faltblattes [13]

Innerhalb des β-Faltblattes verlaufen die β-strands parallel und bilden intersequenzielle Wasserstoffbrücken aus, die das Strukturelement stabilisieren. Ein β-strand ist für gewöhnlich 3 bis 10 Aminosäuren lang.

Die Strukturvielfalt, die sich aus β-Faltblättern ergeben kann ist mindestens ebenso mannigfaltig wie die der α-Helices. Ein Beispiel hierfür sind riesige tonnen- oder porenförmige Strukturen – sogenannte β-Barrels, bei denen mehrere im Kreis gewundene β-Faltblätter einen Zylinder ergeben, der bis zu mehrere Dutzend Ångström im Durchmesser betragen kann. Die Aufgabe solcher Proteinstrukturen liegt meist im Transport anderer Moleküle.

Insgesamt lassen sich also Sekundärstrukturelemente im Wesentlichen über Winkelkonformationen der primärstrukturellen Polypeptidkette klassifizieren. Eine günstige Visualisierung dieser Verhältnisse bietet der Sasisekharan-Ramakrishnan-Ramachandran-Plot [14].

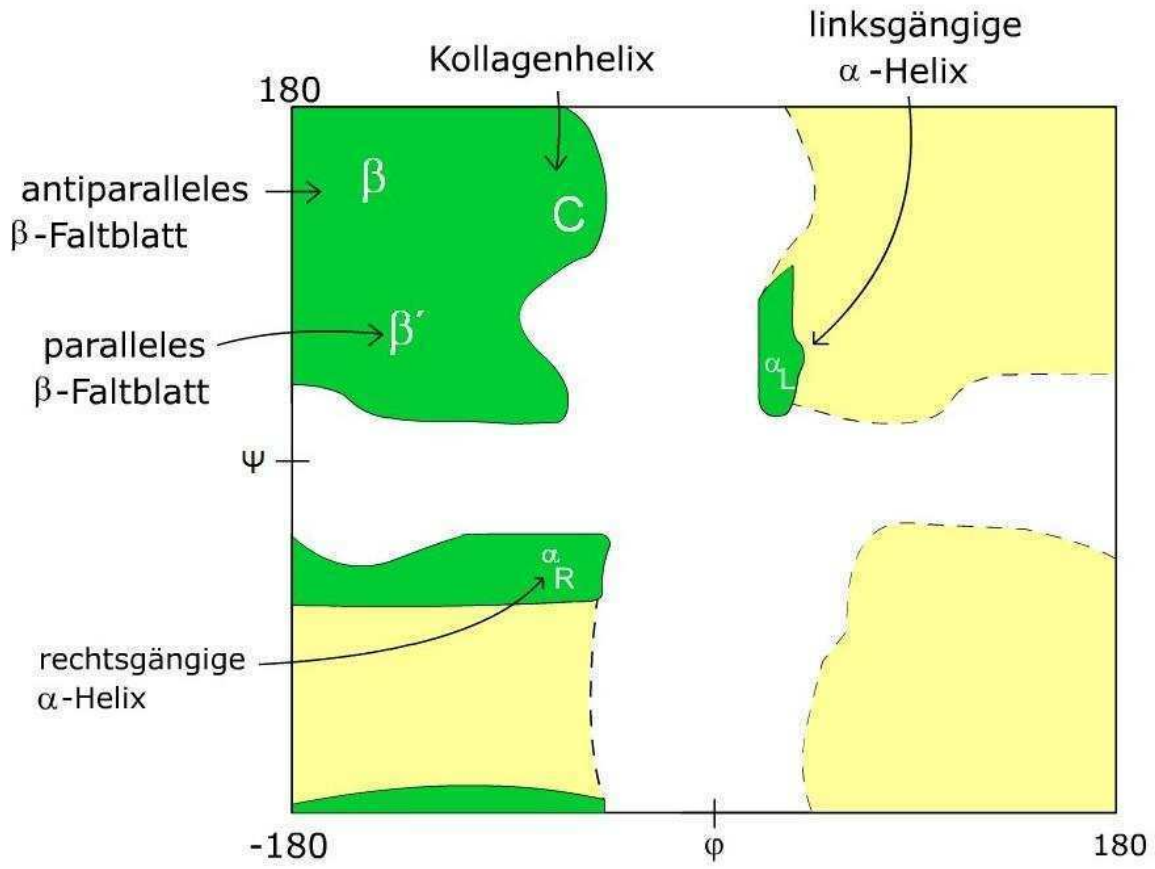


Abbildung 9: Ramachandran-Plot [15]

Im Ramachandran-Plot sind die φ- und ψ-Torsionswinkel eines Proteins gegeneinander abgetragen.

Was im Ramachandran-Plot ersichtlich ist, ist dass sich die Sekundärstrukturelemente α-Helix und β-Faltblatt nur in bestimmten Winkelbereichen der φ- und ψ-Torsionswinkel vorkommen. Dies bedeutet, dass nur Peptidketten, die bestimmte Winkelkonformationen einnehmen charakteristische Sekundärstrukturelemente ausbilden.

Die linksgängige α-Helix kommt in der Natur selbst nicht vor. Was in obenstehender Abbildung mit C gekennzeichnet ist, ist die sogenannte Kollagenhelix. Sie stellt eine weitere Besonderheit unter den helikalen Strukturen dar, da sie aus drei umeinander gewundenen Polypeptidsträngen besteht [16].

Alle Bereiche außerhalb der grün markierten Sekundärstrukturbereiche werden als Random Coil-Bereiche bezeichnet und besitzen eine eher undefinierte Struktur, obgleich auch sie diverse Funktionen im Protein möglich machen.

Sie ermöglichen beispielsweise Rückwärtsbiegungen der Primärstruktur und erlauben dadurch die Ausbildung von sehr kompakten Proteinstrukturen. Ferner agieren sie als eine Art von „Scharnieren“, die es Proteinuntereinheiten vor allem bei Enzymen erlauben, sich gegeneinander zu verschieben und so mögliche Liganden einzuschließen oder sich mit anderen Proteinen zu assoziieren.

Häufig ist an den Enden von Helices oder Faltblättern – an den Stellen, an denen sie in einen Coil-Bereich übergehen – die Aminosäure Prolin zu finden. Dies ist dadurch bedingt, dass Prolin aufgrund seiner ringförmigen Struktur in keines der beiden Sekundärstrukturelemente passt und somit als Strukturbrecher gilt.

Statistisch gesehen teilen sich die Sekundärstrukturelemente in Proteinen wie folgt auf: (Helices – 32%, Faltblätter – 21%, Coils – 47%). Eine solche Verteilung und vor allem die Präferenz jeder Aminosäure sich in einem bestimmten Strukturelement aufzuhalten sind wichtige Grundlagen für Sekundärstrukturvorhersagealgorithmen wie beispielsweise den Chou-Fasman oder den GOR-Algorithmus. Derartige Methoden, die allein aus statistischen Daten abgeleitete Vorhersagen treffen, liegen in ihrer Vorhersagegenauigkeit allerdings nur bei etwa 50-65% [17], [18]. Höhere Methoden der Künstlichen Intelligenz wie Neuronale Netze oder Hidden-Markov Modelle liefern auf diesem Feld bessere Vorhersagegenauigkeiten von etwa 70-80% [19].

Ein wichtiges Ziel der genannten Sekundärstrukturvorhersagemethoden ist es einen Zusammenhang zwischen Sequenz eines Proteins und seiner Faltung herzustellen. Ist die Anordnung der Sekundärstrukturelemente entlang der Primärstruktur bekannt, so lässt sich darauf schließen, wie diese im Raum angeordnet sind, was wiederrum ein wichtiger Anhaltspunkt für einen möglichen Aufbau der gesamten dreidimensionalen Struktur eines Proteins ist.

2.2.3 Tertiärstruktur

Die nächste und eine der wichtigsten Abstraktionsebenen für Proteine ist die Tertiärstruktur. Sie beschreibt die dreidimensionale Anordnung aller Atome, aus denen das Protein aufgebaut ist im Raum und wie diese miteinander verbunden sind. Demzufolge ist dies die erste Abstraktionsebene, die auch die Anordnung der Sekundärstrukturelemente im Raum beschreibt.

Durch den Prozess der Proteinfaltung (auf den im nächsten Abschnitt eingegangen werden soll) entsteht aus der Sekundärstruktur die Tertiärstruktur, und damit die Strukturebene, in der Proteine biologisch erst wirksam werden. Bei globulären (im Cytoplasma vorkommenden) Proteinen ist die treibende Kraft der Tertiärstrukturbildung der sog. hydrophobe Effekt. Dabei werden die Teilbereiche der Peptidkette so angeordnet, dass hydrophobe Bereiche im Proteininneren verborgen werden und hydrophile Bereiche dem umgebenden wässrigen Milieu zugewandt werden. In der Stabilisierung von Tertiärstrukturen spielen weiterhin diverse kovalente und nicht-kovalente Bindungen eine entscheidende Rolle: Disulfidbrücken (stärkste Bindung), Ionenbindungen, Wasserstoffbrücken und hydrophobe Wechselwirkungen (schwächste Bindung).

Die Aminosäure Cystein besitzt ein stereochemisch zugängliches Schwefelatom, welches mit dem eines anderen Cysteinmoleküls eine kovalente Bindung innerhalb der Proteinstruktur eingehen kann.

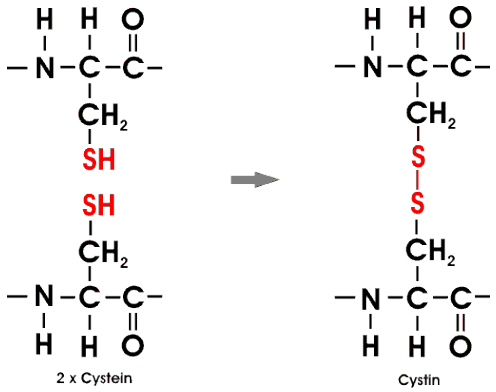


Abbildung 10: Disulfidbrücke zweier Cysteine [20]

Unter Abspaltung zweier Wasserstoffatome bildet sich zwischen den Schwefelatomen (S) zweier Cysteinmoleküle eine stabile Atombindung.

Disulfidbrücken formen und stabilisieren die dreidimensionale Struktur des Proteins oder verknüpfen mehrere Aminosäureketten zu einem funktionsfähigen Protein. Folglich kann das Trennen von Disulfidbrücken zum Funktionsverlust eines Proteins führen.

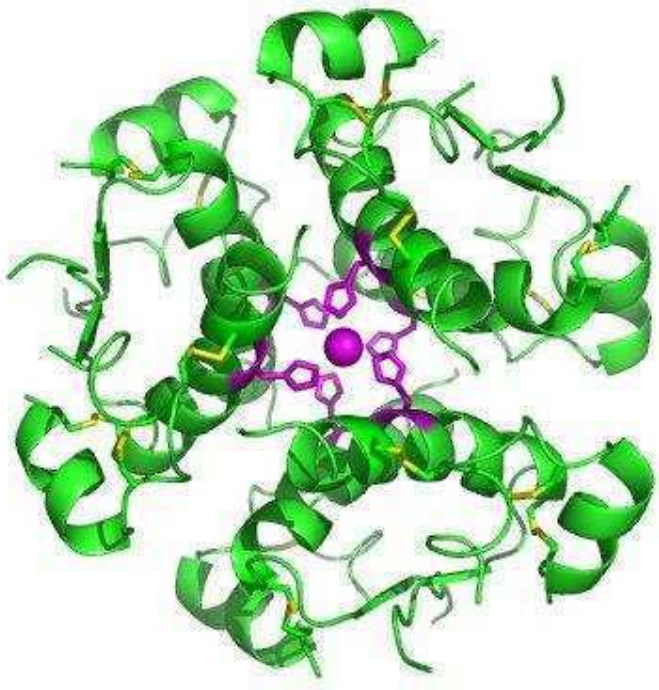


Abbildung 11: Insulin-Hexamer [21]

Für den Stoffwechsel essentielle Proteine wie das Insulin werden durch drei Disulfidbrücken (gelb dargestellt) stabilisiert, wodurch diese ein Hexamer aufbauen und über sechs Histidine Zink-Ionen ligieren können. Beim Abbau des Insulins werden die Disulfidbrücken gespalten, wodurch es wirkungslos wird.

Nicht-kovalente Bindungen wie Ionen- und Wasserstoffbrückenbindungen oder hydrophobe Wechselwirkungen tragen ebenfalls zur Stabilität der Tertiärstruktur bei, jedoch in geringerem Maße. Beispielsweise wird durch ionische Wechselwirkung im Innern eines Insulin-Hexamers ein Zink-Atom ligiert, woran mehrere Insulin-Moleküle beteiligt sind. Diese Bindung bewirkt, dass mehrere Insulinmoleküle als kompakter Proteinkomplex (Zink-Insulin-Komplex) gespeichert werden können.

Eine solche Zusammenlagerung mehrerer Tertiärstrukturen wird als Quartärstruktur bezeichnet und stellt die letzte Abstraktionsebene der Proteine auf struktureller Ebene dar.

2.3 Proteinfaltung

Betrachtet man eine Polypeptidkette mit einer Länge von 100 Aminosäuren und nimmt man an, dass jede Aminosäure in einer von drei Sekundärstrukturkonformationen (Helix, Faltblatt, Coil) befinden kann, so ergeben sich 3^{100} kombinatorische Möglichkeiten die Kette strukturell anzuordnen. Anders ausgedrückt sind dies etwa 10^{48} verschiedene Konformationen. Da die Rotation um eine Bindung in der Polypeptidkette höchstens 10^{14} mal pro Sekunde geschieht lässt sich ausrechnen, dass die zufällige Suche eines Proteins nach seiner natürlichen und stabilen Struktur etwa 10^{26} Jahre dauern würde.

Das Alter des Universums wird auf etwa 13,75 Mrd. bzw. $10^{10,138}$ Jahre geschätzt, was rein rechnerisch bedeuten würde, dass sich seit dem Urknall kaum mehr als ein Protein vollständig gefalten hätte. Allein die Tatsache, dass diese Arbeit in diesem Moment gelesen wird beweist das Gegenteil. Tatsächlich besitzen die meisten Proteine ohnehin nur eine Halbwertszeit von wenigen Stunden bis Tagen und die physiologisch gefaltete Form erreicht die Masse der Proteine in wenigen Sekundenbruchteilen.

Da der Faltungsprozess von Proteinen also nicht durch zufällige Fluktuationen im theoretisch möglichen Konformationsraum der Peptidkette erklärt werden kann, musste es andere Erklärungen für diesen Prozess geben. Dieser als Levinthal-Paradoxon bekannt gewordene Sachverhalt wurde 1969 vom amerikanischen Molekularbiologen Cyrus Levinthal erstmals formuliert.

Grundlegend ist die komplexe dreidimensionale Struktur eines Proteins bereits in seiner Primärstruktur – seiner Aminosäuresequenz – determiniert. Die physikochemischen Eigenschaften der 20 kanonischen Aminosäuren bedingen, dass nicht jede Aminosäure die gleiche Affinität zu den anderen besitzt. Beispielsweise stoßen sich Aminosäuren gleicher Ladung im Protein ab, während sich solche mit unterschiedlicher Ladung eher anziehen. Es sind folglich nur Konformationen möglich, die energetisch auch stabil sind. Da sich ein Protein noch während es am Ribosom synthetisiert wird bereits faltet und analog zu den verschiedenen Strukturebenen, verschiedene Ebenen des Faltungsprozesses unterscheiden lassen lässt sich der Faltungsprozess als Weg durch eine Energielandschaft beschreiben.

Dabei kann zwischen rauen und flachen Energielandschaften unterschieden werden, wobei der Übergang zwischen beiden fließend ist.

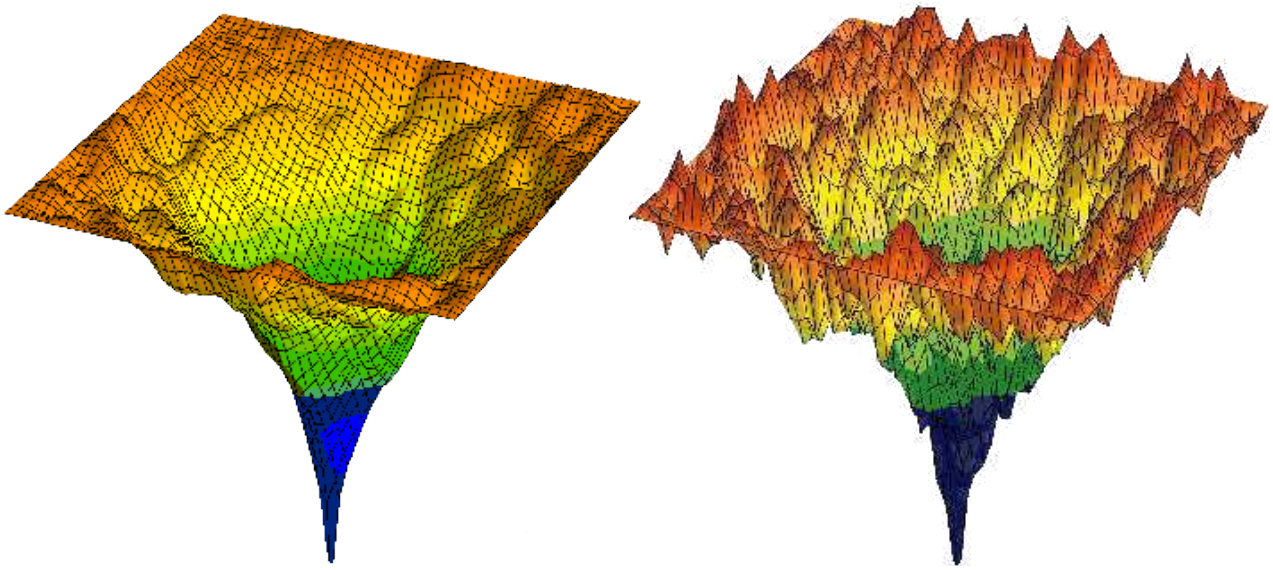


Abbildung 12: Flache und Raue Energielandschaft [22]

In der Energielandschaft sind auf x- und y- Ebene die verschiedensten konformationellen Möglichkeiten (Konformationsraum) der Peptidkette beschrieben. Auf der z-Ebene sind durch Farben die energetischen Zustände der Konformationen beschrieben, wobei blau den niederenergetischsten und damit natürlichsten Zustand darstellt.

Der Konformationsraum, der in der Energielandschaft in der Ebene abgetragen ist, hängt eng mit der thermodynamischen Größe der Entropie, also der Anzahl der möglichen Konformationen, die eine Polypeptidkette einnehmen kann, zusammen. In den Raum ragt aus der Ebene die freie Energie G, die beschreibt ob eine Konformation energetisch stabil ist.

Besonders kleine Proteine wie Ribonucleasen falten sich sehr schnell in ihre korrekte native Struktur, da sie im Gegensatz zu größeren Proteinen schneller ihr Energieminimum und damit ihre natürliche und funktionell aktive Struktur finden. Das Energieminimum kann als globales Minimum der Faltungslandschaft verstanden werden. Große Proteine, vor allem solche, die sich aus vielen Untereinheiten zusammensetzen finden hingegen langsamer ihr Energieminimum. Es ist möglich, dass sie in lokale Energieminima fallen und es somit zu unvollständig gefalteten und funktionell unwirksamen Proteinstrukturen kommen kann.

An dieser Stelle können Hilfsproteine –sogenannte Chaperone wie etwas das Hitzeschockprotein Hsp90 – Abhilfe schaffen. Sie beschleunigen die korrekte Faltung und Assoziation der Proteine, ohne dabei selbst Teil der Struktur zu werden, was impliziert, dass nur nicht-kovalente Bindungen vom Chaperon indirekt geknüpft werden.

Der am besten studierte Chaperon-Mechanismus ist der der Gruppe Hsp60. Die Proteine dieser Gruppe können als Fass-ähnliche Strukturen beschrieben werden, die das zu faltende Protein in sich aufnehmen. Dabei sind an der Innenseite des Fasses hydrophobe Aminosäureketten lokalisiert, die mit den darin befindlichen hydrophoben Ketten des ungefaltenen Proteins in Wechselwirkung treten [23]. Dadurch wird verhindert, dass die hydrophoben Bereiche des Zielproteins einer unerwünschten Aggregation unterliegen. Ist das Proteininnere des Zielproteins korrekt gefalten sind die hydrophoben Außenketten selbst abgesättigt und das Protein wird, nachdem sich die Deckel des Fasses geöffnet haben, das Protein aus dem Chaperon freigesetzt [23].

Außerdem spielen Chaperone bei der Ent- und Refaltung von Proteinen eine Rolle, da diese die Biomembran nur als langgestreckte Polypeptidkette passieren können. Bei diesen Prozessen der Faltung werden enorme Mengen ATP verbraucht. Die Chaperone selbst sind auf genomischer Ebene in hochkonservierten Bereichen codiert.

Um zu erklären, wie Proteine ihren nativen Zustand erreichen existiert eine Vielzahl an Modellen. Drei von ihnen sollen im Folgenden kurz beschrieben werden. Im Gerüstmodell wird angenommen, dass sich als erstes Sekundärstrukturelemente unabhängig von der Tertiärstruktur bilden. Die gebildeten Strukturelemente setzen dann die dichtgepackte Tertiärstruktur zusammen, wobei Bindungen wie Disulfidbrücken gebildet werden, die der Struktur ihren schlussendlichen Charakter geben. Dieses Modell kann vor allem die Faltung kleiner Proteine geringer Strukturkomplexität erklären. Das Modell des Hydrophoben Kollapses geht davon aus, dass die Primärstruktur zu Beginn der Faltung einem relativ gleichmäßigen Kollaps des Proteinmoleküls unterliegt. Das heißt, dass Kettenteile mit hydrophoben Eigenschaften ins Proteininnere gedrängt werden und solche mit hydrophilen Eigenschaften nach in äußere Bereiche driften. Nach diesem Kollaps liegt die Polypeptidkette bereits ähnlich der Form der späteren Tertiärstruktur vor. Gebildet werden nun noch die Sekundärstrukturelemente, die der Struktur ihr endgültiges Aussehen verleihen.

Ein weiteres Modell ist das des Nukleations-Kondensationsmechanismus, in welchem davon ausgegangen wird, dass die Bildung eines frühen diffusen Faltungskeims die weitere Faltung katalysiert. Es stellt ein Analogon zum Prozess der Keimbildung in der Thermodynamik dar. Beim Sieden von Wasser verhält es sich beispielsweise so, dass die Flüssigkeit nicht an allen Stellen gleichzeitig und homogen verteilt in die gasförmige Phase übergeht. Es gibt kleine lokale Zentren, sogenannte Nukleationsstellen, die als erste eine gewisse Energiebarriere überwinden, die sich aufbaut, bevor erste gasförmige Tröpfchen gebildet werden. Diese befähigen ihre Umgebung dazu sich mit bereits vorhandenen Bereichen der neuen stabilen Phase zu aggregieren. Ähnlich soll sich auch Proteinfaltung beschreiben lassen: Der Keim besteht hauptsächlich aus einigen benachbarten Aminosäuren, die bereits eine korrekte Sekundärstrukturwechselwirkung ausgebildet haben. Dieser Keim zieht den Rest der noch nicht vollständig gefaltenen Peptidkette immer weiter in Richtung des globalen Energieminimums. Vom Faltungskeim aus assemblieren sich die restlichen Bereiche der Kette zur vollständigen Tertiärstruktur.

Alle drei Modelle können auf Proteine, die metastabile Zwischenzustände (Intermediate) besitzen, erweitert werden. Den letzten Schritt der Faltung stellt normalerweise das Einrasten der Seitenkette in ihre spezifische Konformation dar. Obwohl jedes Modell eine Erklärung der Faltung für sich bietet, muss angenommen werden, dass diese Prozesse teilweise parallel ablaufen und sich gegenseitig forcieren, um zur endgültigen Tertiärstruktur zu gelangen.

2.4 Proteinmotive

Grundlegend lassen sich Proteinmotive in strukturelle und funktionelle Motive unterscheiden. Datenbanken wie die PDBeMotif oder die Prosite stellen Informationen darüber bereit, wo sich bestimmte Motive in bestimmten Proteinstrukturen befinden. Diese Informationen sind dabei im Falle der Strukturmotive aus Winkelkonformationen und im Falle der funktionellen Motive aus Sequenzmustern abgeleitet. Das heißt, nicht jedes Motiv, dass auf den entsprechenden Datenbanken als Motiv annotiert ist muss auch tatsächlich in der Struktur eine biologisch relevante Bedeutung haben. Dennoch lassen sich natürlich statistische Charakteristika aus Datensätzen von Motiven ableiten.

2.4.1 Strukturelle Motive

Strukturelle Motive sind kurze - etwa drei bis sechs Aminosäuren lange - Strukturfragmente, die meist unabhängig von sequentiellen Eigenschaften definiert sind. Vielmehr spielen ϕ- und ψ-Winkel, sowie Wasserstoffbrückenbindungen eine Rolle. Ein strukturelles Motiv wird über die Anzahl und die Lokalisation von Wasserstoffbrücken, sowie eine definierte Abfolge von Bindungswinkeln als solches annotiert. Sie können in 13 klar voneinander abgrenzbare Gruppen eingeteilt werden: Alphabeta-motif, Asx-motif, Asx-turn, Betabulge, Betabulge-loop, Beta-turn, Gamma-turn, Nest, Niche, Schellmann-loop, St-motif, St-staple, St-turn [2]. Diese 13 Typen lassen sich meist noch in kleinere Subtypen aufspalten, wie etwa Betabulge-loops der Länge 5 oder 6, oder inverse und nicht-inverse Gamma-turns, um nur einige wenige Beispiele zu nennen (Für nähere Informationen siehe Anhang 1 – Erläuterungen zu den Strukturmotiven). Sie finden Verwendung in der Vorhersage von Sekundär- oder Tertiärstrukturen [24] oder im Molekulardesign, wenn Motive auf aktive Stellen bzw. Bindungsstellen für Liganden abgebildet werden [25]. Bis auf einige Ausnahmen (Asx- und St-Motive), bei denen die erste Position im Motiv durch eine definierte Aminosäure festgelegt ist, sind alle Motive rein strukturell über eine Abfolge von Winkeln definiert. Jedes Motiv ist demzufolge Strukturfragment und gleichzeitig ein Vektor von Winkeln.

Die Aufklärung von strukturellen Motiven erfolgt also so, dass Strukturen auf charakteristische Winkelvektoren hin untersucht werden. Weist ein Strukturbereich einen vorher definierten Winkelvektor auf, so wird dieser Bereich einem bestimmten Motivtypus zugeordnet.

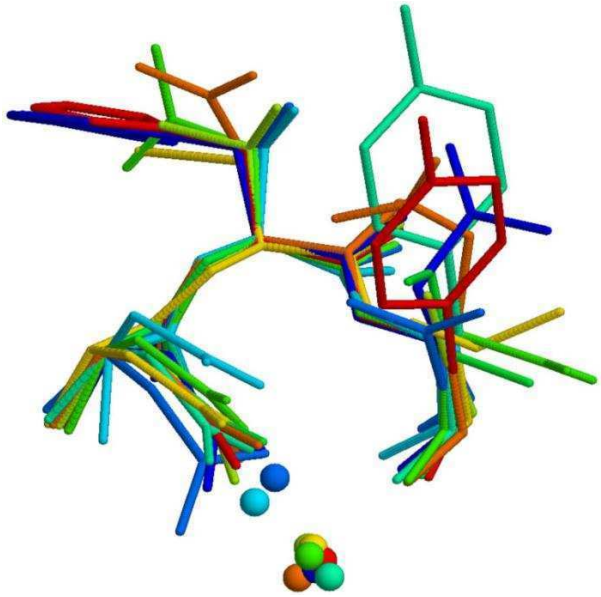


Abbildung 13: Multiples Strukturalignment von Asx-turns [4]

Asx-turns bestehen aus drei Aminosäuren, wobei die erste Position durch Asparagin oder Asparaginsäure realisiert ist. An dieser Stelle sind oftmals Calcium-Ionen koordiniert.

Im multiplen Strukturalignment fällt auf, dass die, auf Grundlage von Winkelkonformationen definierten, Strukturmotive sich strukturell in einem relativ engen Toleranzbereich bewegen. Dennoch können auch kleinere Abweichungen in der Winkelkonformation große Auswirkungen auf die Stabilität des Motivs haben, besonders wenn sich Aminosäuren mit sehr großen Seitenketten – wie die des Tryptophans – im Motiv befinden.

Aus dieser Überlegung lässt sich schlussfolgern, dass sich Motive, je nach sequentieller und damit struktureller Ausprägung, in verschiedene Gruppen einordnen lassen.

Die Tatsache, dass die betrachteten Strukturmotive sequentiell wenig Informationsgehalt besitzen kann mit einem Weblogo [26] visualisiert werden:

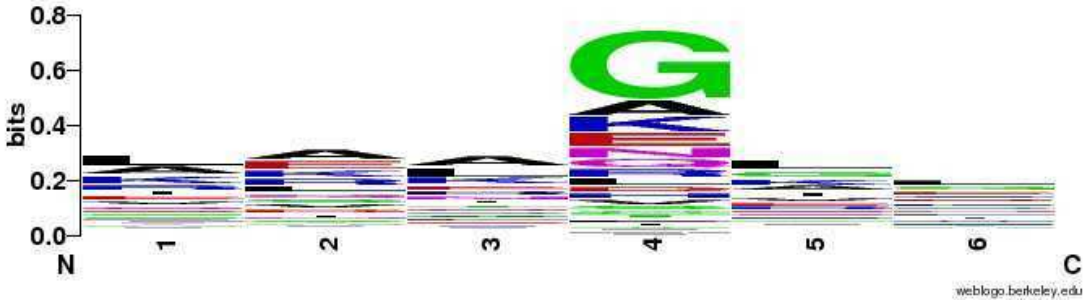


Abbildung 14: Weblogo für Schellmannloops [26]

Der positionsweise sequentielle Informationsgehalt von 542 nicht-redundanten Schellmannloops zeigt, dass sich die Strukturmotive auf Sequenzebene, bis auf Glycin (G) sehr ambivalent verhalten.

Eine besondere Rolle spielt die Aminosäure Glycin in strukturellen Motiven: Aufgrund ihrer fehlenden Seitenkette wird sie meist in das Innere eines Motivs gedrängt und wirkt dort als seitenkettenloses strukturelles Bindeglied. Eine weitere Aufgabe besitzt das Glycin in Nest-Motiven, die oftmals an der Bindung von Phosphaten oder Eisen-Sulfat-Komplexen im Protein beteiligt sind. Da derartige Verbindungen vor allem in frühen Stadien der Erdgeschichte für Metabolismen eine Rolle spielten ist das Nest-Motiv ein möglicher Kandidat für eines der ersten primordialen Proteinstrukturmotive [27].

2.4.2 Funktionelle Motive

Anders als strukturelle Motive sind funktionelle Motive vor allem aus sequentiellen Charakteristika abgeleitet. Ein funktionelles Motiv ist eine Stelle in einer Proteinstruktur, die eine biologische Funktion hat. Dies wären beispielsweise katalystische Stellen von Enzymen oder Stellen im Protein, die dafür verantwortlich sind Moleküle wie ATP, GTP, DNA oder ganze andere Proteine zu binden. Jedes funktionelle Motiv besitzt ein Pattern (engl. für Muster), welches seine sequentielle Ausprägung beschreibt (Siehe auch Anhang Teil 2 - Erläuterungen zu funktionellen Motiven).

Derartige Motive lassen sich durch den Abgleich bekannter (experimentell ermittelter) aktiver Stellen von Proteinen mit gemeinsamer Funktion bestimmen [28]. Sind beispielsweise drei bekannte aktive Stellen gleicher Funktion in verschiedenen Proteinen identifiziert wurden lässt sich aus deren Sequenzen ein mögliches Pattern für ein funktionelles Motiv ableiten:

Sequenz A – ALRDFATHDDF

Sequenz B – SMTAEATHDSI

Sequenz C – ECDQAATHEAS

Ein gemeinsames Muster wäre zwischen diesen drei Sequenzen das vierstellige Pattern A-T-H-[D oder E]. Für die Definition eines Patterns müssen allerdings oftmals weitere Metadaten, wie physikochemische Eigenschaften der Aminosäuren herangezogen werden, um signifikante Muster zu identifizieren und eventuelle falsch-positive Ergebnisse auszuschließen [28].

Ist ein Pattern gefunden, so kann sein Informationsgehalt über einen Datensatz gesehen durch ein Weblogo [26] visualisiert werden:

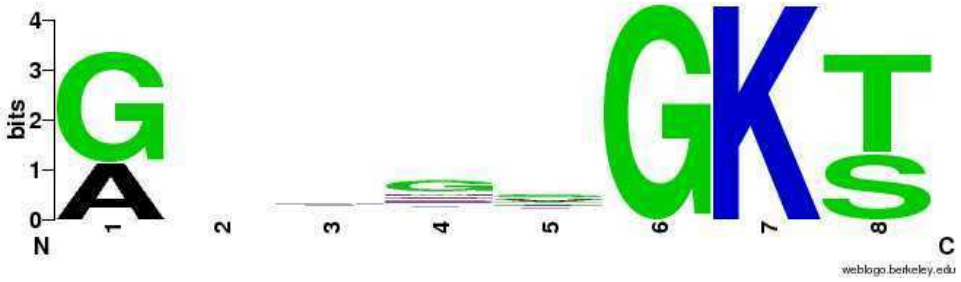


Abbildung 15: Weblogo für PS00017 [26]

Anders als strukturelle Motive sind funktionelle Motive aus sequentiellen Charakteristika abgeleitet und besitzen dadurch auf dieser Ebene einen höheren Informationsgehalt.

Im obenstehenden Weblogo sind 198 Ausprägungen einer ATP/GTP-Bindungsstelle (Prosite-ID: PS00017) auf ihren sequentiellen Informationsgehalt hin überprüft wurden. PS00017 ist acht Aminosäuren lang und leitet sich aus dem Pattern [AG]-x(4)-G-K-[ST] her. An erster Position befindet sich stets ein Alanin (A) oder ein Glycin (G), währen die nächsten vier Positionen durch jede beliebige (x) Aminosäure realisiert sein können, was den sehr geringen Informationsgehalt an diesen Positionen erklärt. Position 6 und 7 sind stets durch Glycin bzw. Lysin (K) realisiert. Die achte und letzte Position wird ähnlich wie die Erste zu gleichen Teilen durch die Aminosäuren Serin (S) oder Threonin (T) besetzt.

Was funktionelle Motive besonders interessant macht, ist die Tatsache, dass sie untereinander strukturell weitaus divergenter sein können als strukturelle Motive. Dies ist besonders im Hinblick darauf interessant, dass funktionelle Motive gleicher zwar eine gemeinsame Funktion besitzen, jedoch sequentiell und damit vor allem strukturell viele mögliche Ausprägungen einnehmen können.

2.5 Theorie der Energieprofile

Ein Energieprofil stellt sich formal gesehen als die Transformation der dreidimensionalen Struktur eines Proteins in einen zweidimensionalen Vektor dar. Dieser Vektor ist für jedes Protein einzigartig, wonach es sich um einen eindeutigen Fingerprint handelt. Die Transformation in ein Energieprofil kann mit allen Proteinen vorgenommen werden, von denen eine 3D-Struktur bekannt ist und somit die Koordinaten aller Atome im Protein vorliegen [10].

Den Energieprofilen, die hier betrachtet werden sollen, liegt ein Ansatz aus der statistischen Physik zugrunde. In diesem werden Energien verwendet, die sich aus der Tendenz einer Aminosäurenseitenkette in der Struktur eher nach außen oder innen gerichtet zu sein, berechnen lassen. Betrachtet man diese Tendenz von einem thermodynamischen Standpunkt aus, lassen sich daraus Energien für jede einzelne Aminosäure in der Struktur berechnen [29]. Beispielsweise besitzt ein Cystein, dessen Seitenkette aus der Proteinoberfläche ragt eine höhere Energie als - wie es der Normalfall wäre - ein Cystein, dessen Seitenkette in der Struktur verborgen liegt. Diese Präferenz einer Aminosäurenseitenkette eher nach außen oder eher nach innen gerichtet zu sein hängt zum einen von ihrer Länge und – was viel wichtiger ist - von ihrer physikochemischen Beschaffenheit ab. Entscheidend hierfür sind vor allem Eigenschaften wie Polarität, Ladung oder Hydrophobizität der Seitenkette.

Folglich muss für die Berechnung eines Energieprofils zunächst ein Innen-/Außen-Kriterium eingeführt werden:

$$f(i) = \begin{cases} n_{in,i} + +, & || \, C_\alpha - c \, || < 5 \text{ Å} \ \vee \left(C_{\alpha,i} - C_{\beta,i}\right)\left(C_{\alpha,i} - c\right) < 0 \\ n_{out,i} + +, & \end{cases}$$

(1)

Hierbei ist i eine der 20 kanonischen Aminosäuren, für die die Eigenschaft Innen/Außen ermittelt wird. Der Wert c repräsentiert den Masseschwerpunkt aller C_α-Atome, die sich in einer Kugel mit einem Radius von 5 Å (0,5 Nanometer) um die Aminosäure i befinden. Wie in (1) ersichtlich wird eine Aminosäure als Innen definiert, wenn obige Bedingungen erfüllt sind, andernfalls als Außen.

Wendet man das Innen-/Außen-Kriterium auf jede Aminosäure in der Struktur bzw. in mehreren Strukturen an, lässt sich eine Innen-/Außen-Statistik aufstellen:

Tabelle 2: Innen/Außenverteilung der Aminosäuren [1]

Aminosäure	Innen	Außen	Aminosäure	Innen	Außen
Cys	4582	1016	**His**	6419	3366
Ile	20370	4141	**Gly**	16698	14326
Ser	12576	10411	**Asp**	10001	14327
Gln	7373	7752	**Leu**	30615	7107
Lys	9285	15193	**Arg**	11327	10441
Asn	8225	8928	**Trp**	4001	1193
Pro	9135	9423	**Val**	23562	6551
Thr	12537	9622	**Glu**	11165	18091
Phe	13353	2813	**Thr**	11228	3529
Ala	22725	11052	**Met**	7003	1723

Ersichtlich ist, dass einige Aminosäuren wie Phenylalanin, Isoleucin oder Valin klare Präferenzen bezüglich ihres Innen-/Außen-Verhaltens haben. Andere wie etwa Asparagin oder Glutamin verhalten sich eher ambivalent. Aus dieser Statistik lässt sich über die inverse Boltzmann-Verteilung (2) die Energie e_i der Aminosäure i nach der Proteinfaltung berechnen.

$$e_i = -k_B\, T\, ln(\frac{n_{in,i}}{n_{out,i}}) \qquad (2)$$

Da die Boltzmannkonstante k_B (1,3806488 * 10^{-23} Joule/Kelvin) und die Temperatur T als konstant angenommen werden können vereinfacht sich die Gleichung auf (3).

$$e_i^* = -\ln\left(\frac{n_{in,i}}{n_{out,i}}\right) \qquad (3)$$

Durch diesen Vereinfachungsschritt verändert sich zwar der numerische Wert der Energie, seine Aussagekraft bleibt jedoch weiterhin erhalten. Die Energie der paarweisen Interaktionen der Aminosäure i mit anderen Aminosäuren entspricht der Umgebung (Environment) von i und vor allem der Zusammensetzung dieser Umgebung [1].

Anhand der zuvor erstellten Innen-/Außen-Statistik kann über Gleichung (4) ein Präferenzwert P für die Umgebung der Aminosäure i approximiert werden, der mit der Energie der Interaktionen von i korreliert:

$$P_{k\in Env} = \prod_{k\in Env} p_k = \prod_{k\in Env}\left(\frac{n_{in,k}}{n_{out,k}}\right)$$

$$\ln P_{k\in Env} = \sum_{k\in Env} \ln\left(\frac{n_{in,k}}{n_{out,k}}\right) \qquad (4)$$

Analog zu (3) ist die Umgebungsenergie E_{Env} durch Gleichung (5) definiert:

$$E_{Env} = -\ln P_{k\in Env} \qquad (5)$$

In (4) und (5) beschreibt $P_{k\in Env}$ die Präferenzen der Aminosäuren k sich in der Umgebung der betrachteten Aminosäure i zu befinden. Durch Zusammenführung von (3), (4) und (5) ergibt sich folgender Sachverhalt:

$$E_i = -|Env|\, ln\left(\frac{n_{in,k}}{n_{out,k}}\right) - \sum_{k\in Env} ln\left(\frac{n_{in,k}}{n_{out,k}}\right) \qquad (6)$$

Die Umgebung selbst ist definiert durch die Kontaktfunktion g(i,j) (7):

$$g(i,j) = \begin{cases} 1 & || C_{\alpha,i} - C_{\alpha,j} || \leq 8\,\text{Å} \\ 0 & \end{cases} \qquad (7)$$

Im Gegensatz zu (1) wird in (7) eine Sphäre mit einem Radius von 8 Å um i herum gewählt [30].

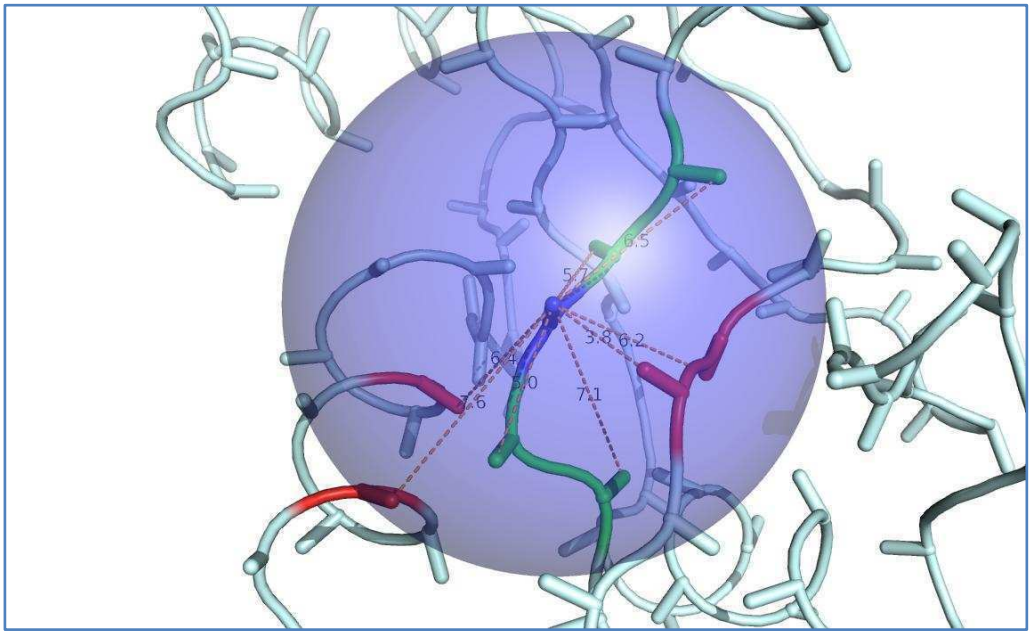


Abbildung 16: 8 Å-Umgebung von His114 in 1B1J [31]

Ein Kontakt zwischen zwei Aminosäuren i und j wird in (7) gezählt, wenn Aminosäure j weniger als 8 Å von i entfernt ist.

Zusammenfassend setzt sich die Gesamtenergie einer Aminosäure i zusammen aus:

$$E_i^* = \sum_{j \in S \setminus i} [g(i,j)\,(e_i^* + e_j^*)] \qquad (8)$$

S ist in (8) als aufgeklärte Proteinstruktur definiert. Durch das Wegfallen von k_B und T in (3) ist die hier betrachtete Energie einheitenlos. Sie ist dennoch direkt proportional zu Energien, die in [J] oder [kcal/mol] notiert sind und dadurch auch mit solchen vergleichbar.

Energieprofile, die mit den soeben gezeigten Methoden berechnet wurden stellen sich grafisch als eine Kurve, die sich im Intervall [10, -50] entlang der Primärstruktur des Proteins bewegt, dar. Dies gilt allerdings nur für globuläre Proteine, da die Berechnung eines Membranproteinenergieprofils auf anderem Wege erfolgt.

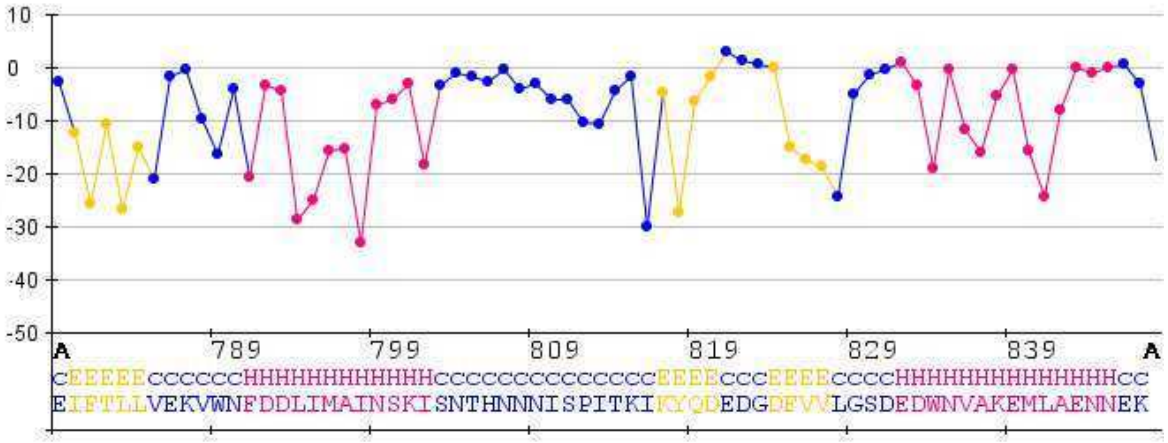


Abbildung 17: Energieprofilausschnitt von 1PQS erzeugt mit eCalc

Im Energieprofil sind besonders die strukturell unschärfer definierten Coil-Regionen (blau) im höher-energetischen Bereich angesiedelt. Stabilere Strukturelemente wie Helices (rosa) und Sheets (gelb) verhalten sich überwiegend alternierend in ihrem Energieverlauf.

Energieprofile können beispielsweise dafür verwendet werden Strukturen mit unbekannter Funktion eine Funktion dadurch zuzuordnen, dass man ihr Energieprofil mit anderen Profilen bekannter Funktion vergleicht. Die ähnlichsten Energieprofiltreffer sind dann mögliche Kandidaten für funktionell-ähnliche Proteine, wodurch Rückschlüsse auf die Funktion der Ausgangsstruktur geschlossen werden können.

In der vorliegenden Arbeit sollen Energieprofile anhand von Motivdaten fragmentiert werden. Die daraus gewonnenen Energievektoren sollen durch Methoden des hierarchischen Clusterings verglichen werden um eventuelle energetische Charakteristika verschiedener struktureller und funktioneller Motive zu extrahieren. Die dafür verwenden Clusteralgorithmen sollen im folgenden Kapitel beschrieben werden.

2.6 Algorithmische Grundlagen

Eine der wichtigsten Aufgaben der Mathematik ist das Auffinden von Mustern bzw. allgemeinen Strukturen in größeren Mengen von Daten jeglicher Art. Was im Folgenden beschrieben werden soll, sind Algorithmen zum Filtern von Mustern in Energieprofilen.

2.6.1 UPGMA

Der UPGMA-Algorithmus (Akronym für „unweighted pair group method with arithmetic mean“) ist ein sehr einfaches und allgemeines hierarchisches Clusterverfahren, dass vielseitige Anwendungen besitzt. Aus einer Menge an Daten, einer vorher auf das Problem zugeschnittenen Distanzmetrik und einer daraus abgeleiteten Distanzmatrix wird ein Distanzbaum konstruiert.

Er wird beispielsweise für das Erstellen phylogenetischer Bäume aus Sequenzdaten genutzt. In diesem Fall ist ein geeignetes Distanzmaß aus Substitutionsmatrizen wie der BLOSUM62 [32] abgeleitet. Im Falle der Energieprofilorientierten Arbeit ist ein solches Distanzmaß anders definiert. Durch die Berechnung eines Energieprofils aus einer Struktur wird diese in einen zweidimensionalen Energievektor transformiert. Strukturfragmente, also strukturelle oder funktionelle Motive stellen sich dann als n-dimensionale Energievektoren dar, wobei n die Länge eines Motivs darstellt.

Über den euklidischen Abstand im n-dimensionalen Raum kann so die euklidische Distanz d zwischen zwei Energievektoren x und y berechnet werden:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (9)$$

Aus diesem einfachen Distanzmaß lässt sich für Energievektoren beliebiger Anzahl eine Distanzmatrix berechnen, die im Anschluss vom UPGMA-Algorithmus genutzt werden können.

Zu Anfang des Verfahrens befindet sich jeder Energievektor in einem eigenen Cluster. In jedem Iterationsschritt werden die beiden Cluster mit dem geringsten Abstand zusammengefasst und die Distanzmatrix neu berechnet. Die Distanz zwischen zwei Clustern ist der Mittelwert der paarweisen Distanzen aller Objekte in beiden Clustern. Angenommen C sei ein neues Cluster, welches aus den Clustern A und B gebildet wurde und es soll geprüft werden, ob ein anderes Cluster D ebenfalls zu C gehört, so muss die Distanz zwischen C und D berechnet werden:

$$d(C,D) := \frac{|A| \times d_{A,D} + |B| \times d_{B,D}}{|A| + |B|} \qquad (10)$$

Ist ein Energievektor einem bestimmten Cluster zugeordnet worden, wird er aus der Distanzmatrix gelöscht. Dieser Vorgang wird so lange iteriert, bis jedes Element einem größeren Cluster zugehörig ist und kein Element mehr in der Distanzmatrix. Das Ergebnis kann als binärer Baum visualisiert werden.

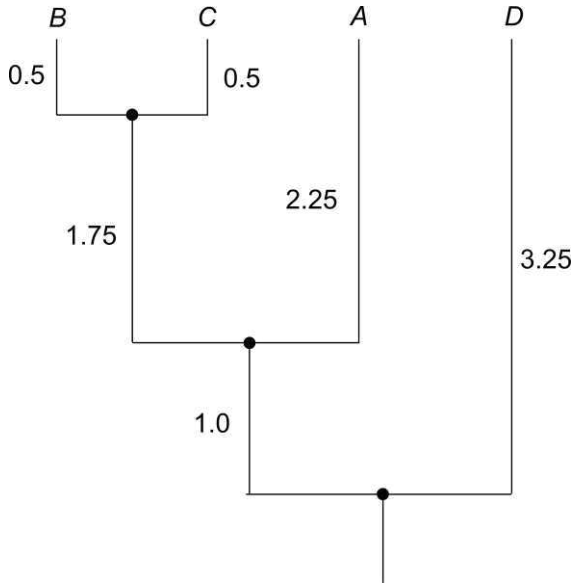


Abbildung 18: UPGMA-Baum [33]

Im binären Baum, den UPGMA als Ergebnis liefert besitzt jedes Blatt den gleichen Abstand zur Wurzel des Baumes. Diese Annahme wird als Molecular Clock bezeichnet und kann in der Phylogenie zu Problemen führen. In den hier betrachteten Anwendungen ist sie jedoch irrelevant.

2.6.2 Neural Gas

Das Neural Gas [34] ist ein künstliches Neuronales Netz, welches sich an die Funktionsweise der Self-Organizing Maps (SOM) anlehnt. Es ist ein Verfahren zur möglichst fehlerfreien Datenkodierung mithilfe von Merkmalsvektoren. Seine Bezeichnung gründet sich auf der Dynamik der Merkmalsvektoren, welche sich während des Lernprozesses wie ein Gas im Datenraum ausbreiten. Es wird neben der Mustererkennung und der Clusteranalyse ebenfalls im Bereich der Spracherkennung [35] oder der Bildkompression [36] eingesetzt.

Neural Gas zählt zu den Vektorquantisierern, das heißt, es ist so konzipiert, dass es anhand einer Trainingsdatenmenge Partitionen im Datenraum erkennt, diese über Prototyp-Vektoren repräsentiert und neue Datenvektoren zu den antrainierten Partitionen zuordnet. Formal werden die einzelnen Komponenten wie folgt bezeichnet:

$V, V \in \mathbb{R}^n$ - der vorliegende Vektor- bzw. Datenraum

$X, X \subset V$ - die Menge der vorliegenden Datenvektoren im Datenraum

$x \in X, x \in \mathbb{R}^n$ - ein Datenvektor mit n Dimensionen (Im Fall des euklidischen Raumes ist $n = 3$, im Fall von Energievektoren ist n die Länge eines Vektors)

$W, W \subset V$ - die Menge der Prototypen

$w, w \in W$ - ein Prototyp

D - ein Abstandsmaß (-metrik)

Zu Anfang wird die Anzahl der Prototypen k, sowie die maximale Iterationszahl t vom Benutzer festgelegt. Da die Anzahl der initialisierten Prototypen die spätere Anzahl der Cluster repräsentiert, muss das optimale Clustering oft durch Probieren gefunden werden. Die Wahl der Anzahl der Iterationen gestaltet sich dagegen praktischer, da mehr Iterationen ein langsameres „Abkühlen des Gases“ und damit eine bessere Konvergenz bedeuten. Weiterhin ist zu Anfang die maximale Lernrate $n_{t_{max}}$ festzulegen. Eine zu hohe Lernrate bedeutet einen chaotischen Lernprozess, eine zu niedrige einen zu langsamen.

Je Iterationsschritt i wird ein Datenvektor x_i zufällig gewählt und die Rangfolge der Prototypen X anhand der einzelnen Abstände $D(x_i, w_i)$ festgelegt. Der Abstand eines Datenvektors zu einem Prototyp ist dabei – analog zum UPGMA-Algorithmus – der euklidische Abstand zweier Vektoren. Der nächstliegende Prototyp zum Datenvektor x_i erhält den Rang 1, der zweitnächste den Rang 2, usw. Die Prototypen müssen also in jedem Iterationsschritt sortiert werden, was eine höhere Rechenzeit mit sich bringt. Die Adaption der Prototypen erfolgt in Abhängigkeit ihres Ranges. Weiterhin spielen die Abstandsparameter λ und $\lambda_{t_{max}}$, die den Rang eines Prototyps relativieren, die bereits bekannte Lernrate n sowie die zuvor definierte Lernrate $n_{t_{max}}$ eine Rolle bei der Prototypenadaption.

Im Detail stellt sich der Algorithmus wie folgt dar:

$Initialisiere\ k\ Prototypen\ (W)\ zufällig\ im\ Datenraum\ V$

$Initialisere\ Iterationszahl\ t_{max}\ , anfängliche\ Lernrate\ n_0\ , max. Lernrate\ n_{t_{max}}\ , \lambda\ und\ \lambda_{t_{max}}$

$t = 1$

$while\ t < t_{\max}\quad do$:

$$n(t) = n_0\left(\frac{n_{t_{max}}}{n_0}\right)^{\frac{t}{t_{max}}}$$

$$\lambda(t) = \lambda_0\left(\frac{\lambda_{t_{max}}}{\lambda_0}\right)^{\frac{t}{t_{max}}}$$

$$h_{r_{w_i,x_i}} = e^{\frac{-rg\,(w_i,x_i)}{\lambda(t)}}$$

$$Update\colon\ \forall w_i \in W : w_i^{t+1} = w_i^t + n(t)\ h_{r_{w_i,x_i}}(x_i - w_i)$$

$t++$

$end\ while$

Im Gegensatz zu einfacheren Clustermethoden wie dem UPGMA oder dem k-means-Algorithmus ist Neural Gas zwar rechenaufwändiger jedoch gleichzeitig numerisch wesentlich stabiler.

2.6.3 Intelligentes Monte-Carlo Sampling

Beim Monte-Carlo Sampling handelt es sich um Verfahren, die es erlauben zufällige Stichproben aus einer Menge an Elementen zu ziehen. Ein intelligentes Monte-Carlo Sampling stellt einen Spezialfall dieser Methode dar. Die Elemente werden zwar weiterhin zufällig aus einer Menge gewählt, jedoch besitzt nicht jedes Element die gleiche Wahrscheinlichkeit gewählt zu werden. Die zugrunde liegende Wahrscheinlichkeitsverteilung ist also nicht gleichverteilt sondern von diversen, eigens definierbaren Parametern abhängig.

Die Motivation hinter dem Verfahren ist es aus einer großen Datenmenge, die sich aus mehreren ungleich partitionierten Teilmengen zusammensetzt zufällige, aber trotzdem repräsentative Elemente zu wählen. Die Tatsache, dass in einem Monte-Carlo Samplingschritt immer zufällig gewählt wird impliziert, dass zwei oder mehr nacheinander ausgeführte Samplings nahezu niemals dasselbe Ergebnis liefern können. Die Herausforderung besteht darin, dass die Ergebnisse zwar unterschiedlich sein dürfen, aber jedes für sich die Grundgesamtheit genauso gut wie ein anderes Ergebnis repräsentieren muss.

In der vorliegenden Arbeit wird ein Intelligentes Monte-Carlo Sampling dafür verwendet, um sehr große UPGMA-Bäume auf kleinere zu reduzieren. Dabei sollen einerseits Energievektoren zufällig aus den Clustern gewählt werden, andererseits soll die Gesamttopologie des Baums erhalten bleiben. Wie diese Aufgabe bewerkstelligt wird soll im entsprechenden späteren Kapitel erläutert werden.

3 Energetische Analyse der Motive

Die energetischen Analysen der strukturellen und funktionellen Motive mittels Clusteralgorithmen und den dazugehörigen Methoden sollen in diesem Kapitel dargelegt werden. Zunächst soll allerdings beschrieben werden, woher die verschiedenen Motiv- und Energiedaten stammen.

3.1 Überblick über die Datensätze

Der Datensatz für die Untersuchung der strukturellen Motive umfasst etwa 2700 Proteinstrukturen, die auf Grundlage von Motivdaten, die von der PDBeMotif [37] bezogen wurden, auf Motive geparst wurden. Dieses Parsing ergab eine Anzahl von 10547 strukturellen Motiven. Die Energieprofile für die 2700 Proteinstrukturen stammen vom Energy-Profile-Server eProS der Bioinformatics Group Mittweida [38]. Der dort vorhandene Datensatz wurde nach diversen Kriterien mit BlastClust [39] gesampelt: So fielen Einträge, die eine Sequenzähnlichkeit von mehr als 25% besaßen aus dem Datensatz heraus.

Weiterhin wurden alle Einträge die aus in-silico generierten Modellen abgeleitet sind entfernt. Der Datensatz für die Untersuchung der strukturellen Motive umfasst demzufolge ausschließlich nicht-redundante aufgeklärte Proteinstrukturen.

Die Motivdaten von der PDBeMotif enthalten die Indexierung sowie die Sequenzen für strukturelle Motive in Proteinen. Anhand dieser Indices wurden aus jedem der 2700 Energieprofile die entsprechenden Energievektoren herausgeschnitten und in ein dafür entworfenes Format abgelegt:

```
PDB_ID  1DEP_5
MOTIF_SEQ   FRKAF
MOTIF_TYPE  alphabetamotif
EP_SEQ  FRKAF
MOTIF_ENTRY F   4   H   -12.568211114691604 3.Quantil
MOTIF_ENTRY R   5   H   -3.3009814692442228 2.Quantil
MOTIF_ENTRY K   6   H   0.8236813890123544  1.Quantil
MOTIF_ENTRY A   7   H   -9.160790637082783  3.Quantil
MOTIF_ENTRY F   8   H   -16.837108502426943 3.Quantil
```

Abbildung 19: Motiv-Eintrag im Motif-Energy-File

Motif-Energy-Eintrag für ein Alphabeta-Motiv von 1DEP. Um Indexierungsfehler zu vermeiden wurde die Motiv-Sequenz der PDBeMotif Daten mit der des Energieprofils abgeglichen.

Auf diese Weise wurden Files generiert, in denen die 10547 Strukturmotive ihren jeweiligen Energievektoren zugeordnet werden.

Die 10547 betrachteten Strukturmotive verteilen sich wie folgt:

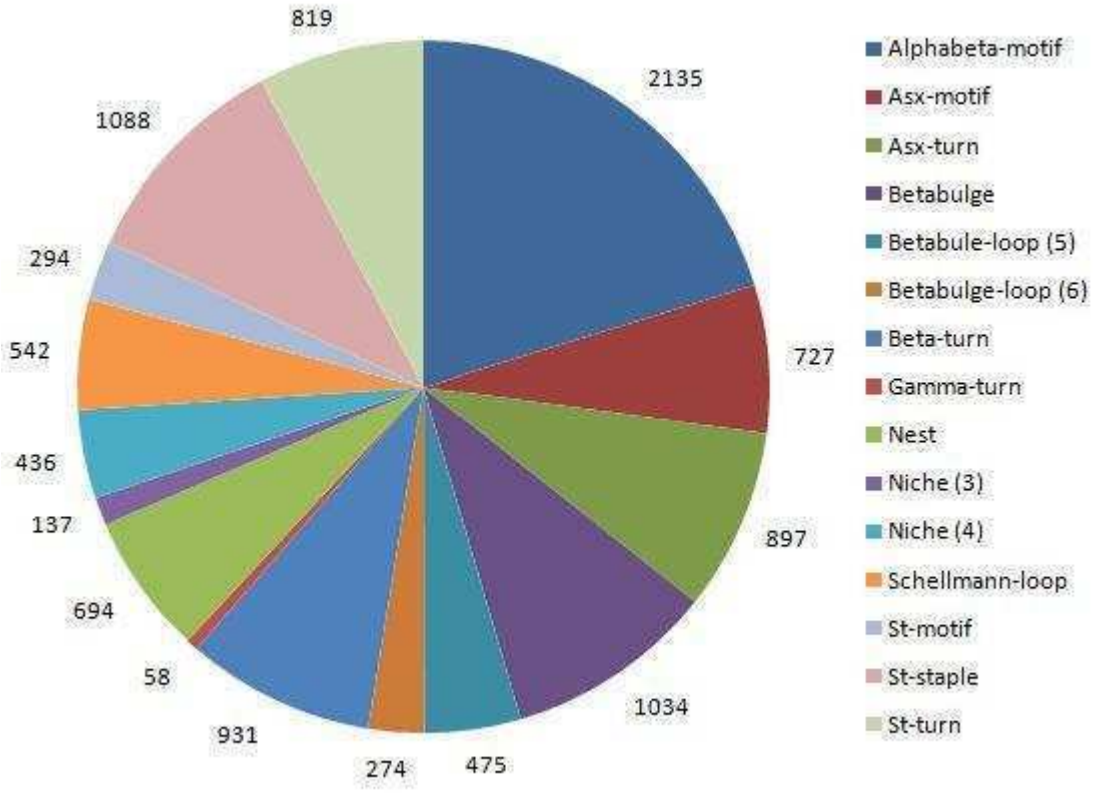


Abbildung 20: Quantitative Verteilung der Strukturmotive

Aufgetragen ist die quantitative Verteilung aller strukturellen Motive im verwendeten Datensatz. Besonders die Motive Schellmann-loop und Betabulge-loop mit einer Länge von 6 Aminosäuren sind eher selten.

Auf die Grundgesamtheit bezogen verhalten sich die strukturellen Motive, bis auf wenige Ausnahmen weitgehend gleichverteilt in ihrem Auftreten. Lediglich Betabulges sind besonders häufig vertreten, was dudurch zu begründen ist, dass der Betabulge erstens nur aus zwei Aminosäuren aufgebaut und zweitens sehr häufig in β-Faltblättern anzutreffen ist.

Der Datensatz, der zur Untersuchung der funktionellen Motive herangezogen wurde besteht aus etwa 5900 Proteinstrukturen. Er wurde mit dem Tool „Representive protein chains from PDB“ [40] aus dem Gesamtdatensatz der PDB gesampelt:

factors	apply constraints	threshold	priority
Resolution	○ No ◉ Yes	X > 3.0 will be eliminated.	1
R-factor	○ No ◉ Yes	X > 0.3 will be eliminated.	2
number of chain break	◉ No ○ Yes	X > 0 will be eliminated.	3
ratio of non-standard residues	◉ No ○ Yes	X > 0 %will be eliminated.	4
ratio of residues with only CA coordinates	◉ No ○ Yes	X > 0 %will be eliminated.	5
ratio of residues with only backbone coordinates	◉ No ○ Yes	X > 0 %will be eliminated.	6
number of residues	◉ No ○ Yes	X < 40 will be eliminated.	7
include MUTANT	◉ No ○ Yes		8
COMPLEX	○ only COMPLEX ◉ exclude COMPLEX ○ All		9
FRAGMENT	○ only FRAGMENT ◉ exclude FRAGMENT ○ All		10
include NMR	○ No ◉ Yes		
include membrane proteins	◉ No ○ Yes		

Abbildung 21: Sampling-Kriterien für funktionellen Motivdatensatz [40]

Anhand der abgebildeten Kriterien wurden 5900 Proteinstrukturen aus der PDB extrahiert und zu Untersuchung der funktionellen Motive herangezogen.

Die in Abbildung 20 aufgelisteten Kriterien schließen beispielweise alle Strukturen mit einer Auflösung von weniger als 3 Å, alle als Komplex vorliegenden Strukturen, alle Membranproteine und alle Strukturen, die lediglich als Fragment vorliegen, aus. Analog zu den strukturellen Motiven wurde auch hier ein Redundanzkriterium eingeführt, dass Strukturen mit einer Sequenzähnlichkeit von mehr als 25 % ausschließt. Aus diesen 5900 Strukturen konnten 5600 Energieprofile berechnet werden.

Diese Differenz von 300 Strukturen resultiert daraus, dass an manchen Proteinstrukturen während der Kristallisation andere kleinere Moleküle „hängengeblieben“ sind. Diese Fremd-Moleküle werden zwar in der Struktur erfasst, können aber nicht vom Algorithmus, der das Energieprofil berechnet verwendet werden. Entsprechend fallen solche Strukturen aus der Berechnung heraus.

Die übrigen 5600 Energieprofile wurden analog zu den strukturellen Motiven mit Motivdaten der Prosite auf funktionelle Motive geparst, was eine Gesamtzahl von 55113 funktionellen Motiven ergab. Auch diesen wurden ihre entsprechenden Energievektoren zugeordnet und in entsprechenden Files abgespeichert.

Die 55113 betrachteten funktionellen Motive sind im Vergleich zu den strukturellen Motiven bezüglich ihres Auftretens wesentlich extremer verteilt:

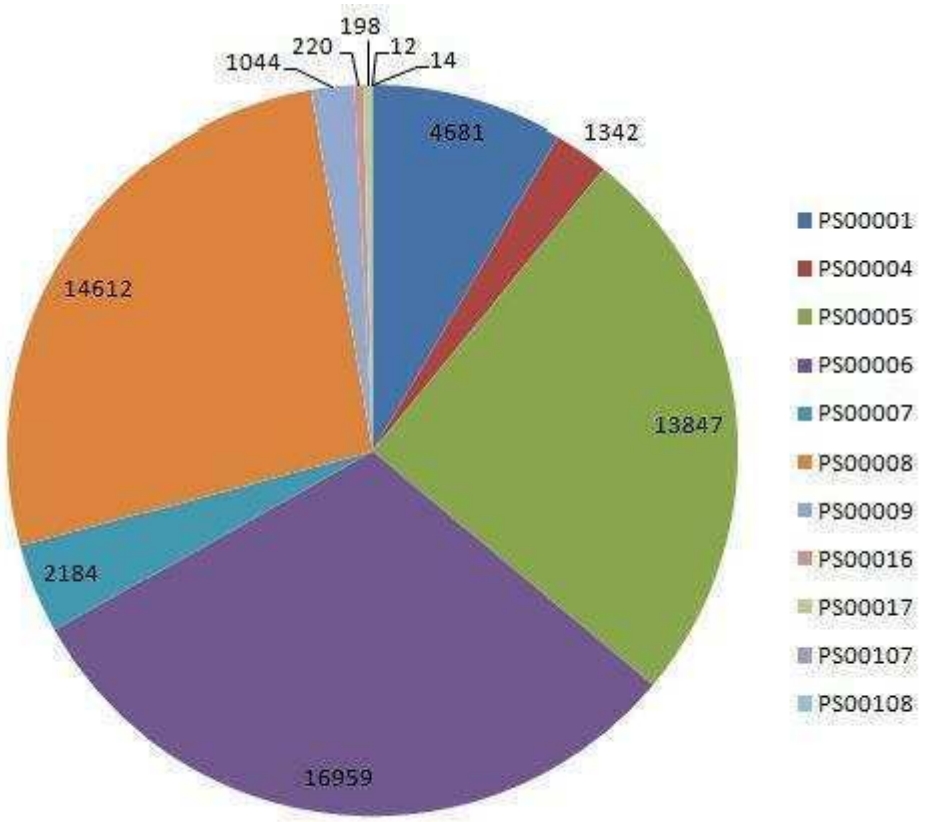


Abbildung 22: Quantitative Verteilung der funktionellen Motive

Trotz der Größe des Datensatzes für funktionelle Motive sind vereinzelte Motive (PS00107 und PS00108) äußerst selten in Proteinen aufzufinden.

Eine Reihe von weiterführenden Informationen zu den funktionellen Motiven, ihren Sequenzmustern und ihrer Funktion findet sich unter Anhang Teil 2 - Erläuterungen zu funktionellen Motiven.

3.2 Erste statistische Analysen

Für die erste statistische Analyse der strukturellen und funktionellen Motive sollen zunächst arithmetische Mittelwerte der Energievektoren berechnet werden. Über die gesamten Datensätze gesehen lassen sich die positionsweisen Energien für jedes Motiv bestimmen und über die Gesamtzahl seiner Ausprägungen mitteln:

Tabelle 3: Mittelwerte der positionsweisen Energien für Strukturmotive

Motif	1.AA	2.AA	3.AA	4.AA	5.AA	6.AA
Alphabeta-motif	-13,3018	-12,9593	-10,5809	-12,8762	-14,8349	
Asx-motif	-2,4108	-8,6609	-5,3766	-9,8333	-16,1001	
Asx-turn	-1,8455	-6,835	-5,4569			
Betabulge	-14,2853	-12,6417				
Betabule-loop (5)	-7,8133	-3,5287	-1,4238	-4,4691	-6,6662	
Betabulge-loop (6)	-6,875	-6,4868	-2,0957	-3,8287	-4,8374	-8,711
Beta-turn	-12,7564	-6,8323	-6,2341	-9,3804		
Gamma-turn	-6,43	-5,0101	-4,9413			
Nest	-5,9415	-9,5233	-11,4304			
Niche (3)	-8,6524	-7,1355	-8,0042			
Niche (4)	-12,087	-6,6186	-5,0061	-12,0449		
Schellmann-loop	-14,1022	-10,3029	-11,7037	-7,7173	-12,9098	-10,9206
St-motif	-5,9024	-9,9543	-5,0016	-6,6248	-18,7908	
St-staple	-13,373	-13,8494	-14,9826	-14,7187	-9,5731	
St-turn	-6,7695	-6,3341	-8,2797			

Aufgeführt sind die positionsweisen Mittelwerte der Energien für alle untersuchten Strukturmotive. Betrachtet wurden dabei alle 10547 Motive aus den etwa 2700 Strukturen.

Den gemittelten Energien für jede Position in den Motiven zufolge gibt es solche Motive, die eher dazu tendieren sich höher-energetisch zu verhalten (Gamma-turns, Asx-turns) und solche, die sich eher in niedrigeren Energiebereichen ansiedeln (Alphabeta-motive, Betabulge).

Eine solche Mittelwertstatistik für positionsweise Energien lässt sich ebenfalls für alle funktionellen Motive im Datensatz erstellen:

Tabelle 4: Mittelwerte der positionsweisen Energien für funktionelle Motive

Motif	1.AA	2.AA	3.AA	4.AA	5.AA	6.AA	7.AA	8.AA	9.AA
PS00001	-9,2462	-12,2997	-10,7909	-12,5308					
PS00004	-7,8488	-8,2492	-11,8382	-9,8429					
PS00005	-9,993	-11,825	-8,7279						
PS00006	-9,9429	-11,5779	-11,0924	-8,2418					
PS00007	-8,041	-11,2551	-11,6831	-10,0255	-10,0925	-12,4964	-13,2292	-13,8282	-15,0974
PS00008	-10,9185	-13,8115	-12,8229	-13,0672	-12,2082	-13,4445			
PS00009	-9,6249	-8,7535	-7,8475	-8,5707					
PS00016	-9,426	-10,2919	-8,9496						
PS00017	-11,348	-9,9597	-10,1308	-8,6826	-10,2141	-9,2779	-7,3816	-10,0964	
PS00107	-14,3221	-9,0555	-8,451	-5,6848	-7,5706	-12,4671	-6,2561	-7,5439	-18,3083
	10.AA	11.AA	12.AA	13.AA	14.AA	15.AA	16.AA	17.AA	18.AA
	-18,8621	-17,8561	-13,8792	-6,3272	-13,7626	-10,2401	-10,5022	-8,1716	-3,7128
	19.AA	20.AA	21.AA	22.AA	23.AA	24.AA			
	-8,5735	-11,2882	-19,3218	-17,6	-21,5276	-8,7483			
	1.AA	2.AA	3.AA	4.AA	5.AA	6.AA	7.AA	8.AA	9.AA
PS00108	-19,0836	-17,2861	-17,1653	-17,3701	-13,0005	-22,9509	-7,1085	-11,1693	-8,3101
	10.AA	11.AA	12.AA	13.AA					
	-6,1526	-17,9037	-21,7563	-18,3137					

Analog zu Tabelle 3 sind hier die positionsweisen Mittelwerte der Energien für alle untersuchten funktionellen Motive. Betrachtet wurden dabei alle 55113 Motive aus den etwa 5600 Strukturen.

In den gemittelten positionsweisen Energien der funktionellen Motive fällt auf, dass einzelne Positionen besonders niedrige Energien einnehmen. Beispielsweise besitzt Position 6 in PS00108 eine niedrige Energie von -22,9509. PS00108 ist durch das Sequenzpattern [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3) definiert. Die sechste Position ist also stets durch die Aminosäuren Leucin, Isoleucin, Valin, Methionin, Phenylalanin oder Tryptophan realisiert. Diese sind alle hydrophobe Aminosäuren und damit physikochemisch sehr ähnlich. Obwohl das Motiv an dieser Stelle also sequentiell recht variabel ist, ist es strukturell sowie energetisch konserviert.

Gleiches gilt in PS00108 auch für die 12. Position, an der dieselben Aminosäuren und mit Cystein und Threonin noch zwei weitere hydrophobe Aminosäuren beteiligt sind. Diese beiden Positionen sind somit die wahrscheinlichsten Kandidaten für die entscheidenden funktionellen Stellen im Motiv, da sie energetisch und damit strukturell stabilisiert vorliegen.

Um die positionsspezifische Verteilung von Sekundärstrukturelementen wie coil, Helix und Strand in einem Motiv zu analysieren lässt sich eine entsprechende Statistik erstellen. Dabei werden die relativen Häufigkeiten der Sekundärstrukturelemente pro Position auf die Gesamtverteilung der Sekundärstrukturelemente normiert.

Sei S_i ein beliebiges Sekundärstrukturelement (coil, Helix, Strand) an der Position i. Und $H_A(S_i)$ dessen absolute Häufigkeit über alle Ausprägungen eines Motivs gesehen. Dann ist $H_R(S_i)$ dessen relative Häufigkeit, die sich errechnet durch:

$$H_R(S_i) = \frac{H_A(S_i)}{n} \qquad (11)$$

n ist die Gesamtanzahl aller Sekundärstrukturelemente an einer Position i. Die relative Häufigkeit eines Strukturelements an einer Position, lässt sich auf die Gesamtverteilung des Strukturelementes über den Gesamtdatensatz gesehen, normieren. Dabei soll $H_R(S_{Ges})$ als die relative Häufigkeit des Strukturelementes S im gesamten Datensatz und $P(S_i)$ der normierte Präferenzwert eines Strukturelementes an der Position i sein:

$$P(S_i) = \frac{H_R(S_i)}{H_R(S_{Ges})} \qquad (12)$$

Auf diese Weise lässt sich für jedes Sekundärstrukturelement seine Präferenz an einer bestimmten Stelle im Motiv zu sein berechnen. Die Gesamthäufigkeiten der Strukturelemente $H_R(S_{Ges})$ sind für beide Datensätze getrennt berechnet worden. Datensatz 1 bezeichnet die Gesamtheit aller Strukturmotive, Datensatz 2 umfasst alle funktionellen Motive.

Tabelle 5: Verteilung der Sekundärstrukturelemente der Datensätze

Datensatz	c	H	S
1	38,33%	55,23%	6,44%
2	32,63%	43,03%	22,33%

Die statistische Verteilung der Sekundärstrukturelemente coil (c), Helix (H) und Strand (S) wird zur Normierung des relativen Auftretens positionsspezifischer Strukturelemente in einem Motiv verwendet.

Für alle Ausprägungen des Betabulgeloop5-Motivs erhält man nach obiger Berechnung die Präferenzen jedes Sekundärstrukturelements sich an einer bestimmten Position im Motiv aufzuhalten:

Tabelle 6: Sekundärstruktupräferenzwerte für alle Betabulgeloop5-Motive

Betabulgeloop (5 AS)				
Position	**c**	**H**	**S**	**Gesamt**
1	1,98	0,00	3,73	5,71
2	2,59	0,00	0,13	2,72
3	2,61	0,00	0,00	2,61
4	2,59	0,00	0,10	2,69
5	2,18	0,00	2,55	4,73

Für 475 Ausprägungen des Betabulgeloop5-Motivs sind die statistischen Präferenzen der Sekundärstrukturelemente ermittelt worden. Präferenzen, die gegen den Wert 1 konvergieren suggerieren ein normales Verhalten des Strukturelementes an dieser Position. Selbiges gilt, wenn die Summer der Präferenzen gegen den Wert 3 konvergiert. In diesem Fall ist die Verteilung der Position als natürlich anzusehen.

Da die Präferenzen für eine Helix (H) an jeder Position gleich Null ist, bedeutet das, dass das Betabulgeloop5-Motiv näherungsweise niemals in einer Helix auftritt. Weiterhin lässt sich aufgrund der hohen Präferenz der Positionen 1 und 5 bezüglich eines Strands (S) und der konstanten Präferenzen an coil-Positionen (c) drauf schließen, dass das Motiv geneigt ist als Brückenelement zwischen Faltblättern zu fungieren. Vergleicht man die Positionen 1 und 5 mit den dazugehörigen gemittelten Energien (vgl. Tabelle 3), so ist ersichtlich, dass diese Positionen wichtig für die Strukturbildung und damit energetisch stabiler als die übrigen sind. Ein sehr ähnliches Verhalten weisen auch andere, mit Faltblättern assozierte Motive wie der Beta-turn oder der Betabulge auf. Sie liegen stets energetisch stabil zwischen Strand-Bereichen, wobei ihr Anfang bzw. ihr Ende auch den Anfang bzw. das Ende eines Strands bilden. Die positionsweisen Sekundärstrukturpräferenzen geben folglich Aufschluss darüber, wie sich ein Motiv in der Struktur verhält, und welche Sekundärstrukturelemente es aufbaut beziehungsweise miteinander verbindet.

Ein weiteres Beispiel dafür, wie die betrachteten Motive Strukturen aufbauen ist das St-Staple-Motiv:

Tabelle 7: Sekundärstruktupräferenzwerte für alle St-Staple-Motive

St-Staple				
Index	**Coil**	**Helix**	**Strand**	**Gesamt**
1	0,19	1,67	0,07	1,93
2	0,12	1,72	0,01	1,86
3	0,1	1,74	0,01	1,85
4	0,12	1,72	0,03	1,87
5	0,27	1,61	0,1	1,98

Sekundärstrukturpräferenzen für 1086 St-Staple-Motive

Die Berechnung der Sekundärstrukturpräferenzen für das St-Staple-Motiv zeigt die eindeutige Neigung des Motivs sich in Helix-Strukturen zu befinden. Ein kaum signifikantes Auftreten zeigt es in coil- beziehungsweise Strand-Bereichen. Das Motiv ist folglich also ein Bauelement für helikale Strukturen in Proteinen. Auch die gemittelten Energien (vgl. Tabelle 3) für alle Positionen eines St-Staple-Motivs zeigen ein niederenergetisches Verhalten, was zeigt, dass die Energien des Motivs direkt mit seinem strukturellen Auftreten korrelieren.

Als Beispiel für coil-typische Motive sind Gamma- und Asx-turns zu nennen:

Tabelle 8: Sekundärstruktupräferenzwerte für Gamma- und Asx-turns

Gammaturn				
Index	**Coil**	**Helix**	**Strand**	**Gesamt**
1	2,38	0,03	1,07	3,49
2	2,61	0	0	2,61
3	2,52	0,06	0	2,58
Asx-turn				
Index	**Coil**	**Helix**	**Strand**	**Gesamt**
1	2,09	0,15	1,8	4,04
2	2,1	0,32	0,27	2,7
3	1,98	0,36	0,63	2,98

Als coil-bildende Motive können Gamma- und Asx-turns gelten, da ihre Präferenzen zu coil-Bereichen geneigt sind.

Vergleicht man erneut die Energien (vgl. Tabelle 3) der Positionen von Gamma- und Asx-turns, so fällt auf, dass sie an allen Stellen einen hohen Wert haben, was verringerte strukturelle Stabilität suggeriert. Erwartungsgemäß verhalten sich dazu auch die Strukturpräferenzen der Motive. Sie liegen zu einem Großteil auf Seiten des coil-Elementes, welches als energetisch eher ungünstiges Bindeglied zwischen größeren und stabileren Strukturelementen zu finden ist.

Auf die Strukturmotive angewandt verraten die strukturellen Präferenzen in Verbindung mit den zugehörigen Energien also, welche Motive welche Sekundärstrukturelemente aufbauen und miteinander verbinden. Die Energien korrelieren dabei klar mit der strukturellen Stabilität der Strukturelemente. Niedrige Energien werden von coil-typischen Motiven wie Gamma, Asx-turns, Nests oder Nichen eingenommen. Im Gegensatz dazu nehmen Motive, die typisch sind stabilere Strukturelemente wie Helices oder Faltblätter zu bilden, höhere Energien ein.

Selbiges Vorgehen, wie soeben gezeigt lässt sich auch für alle funktionellen Motive anwenden. Betrachtet man die Sekundärstrukturpräferenzen für funktionelle Motive und ihre dazugehörigen Energien, so lassen sich daraus Rückschlüsse ziehen, an welchem Strukturelement im Motiv die Funktion lokalisiert ist. Eine niedrige Energie an einer bestimmten Position im Motiv ist ein Hinweis auf strukturelle Konserviertheit dieser Position, was wiederum oft das katalytisch-funktionelle Zentrum von Proteinstrukturen darstellt. Ein besonders gutes Beispiel hierfür ist das Motiv PS00108, eine aktive Stelle von Protein Kinasen:

Tabelle 9: Sekundärstruktupräferenzwerte für alle PS00108-Motive

PS00108 Index	Coil	Helix	Strand	Gesamt
1	0,21	1,00	2,24	3,44
2	0,83	0,83	1,6	3,25
3	1,86	0,33	0,96	3,15
4	1,65	0,33	1,28	3,26
5	1,86	0,17	1,28	3,3
6	1,44	0,17	1,92	3,53
7	1,03	0,83	1,28	3,14
8	1,44	1,00	0,32	2,76
9	1,24	1,00	0,64	2,87
10	1,86	0,33	0,96	3,15
11	1,24	0,33	1,92	3,49
12	0,62	0,5	2,56	3,68
13	0,62	0,5	2,56	3,68

Die Sekundärstrukturpräferenzen von 14 Ausprägungen des PS00108-Motivs zeigen, dass sich funktionelle Motive, bis auf einige besondere Stellen strukturell wesentlich ambivalenter verhalten.

Die Position 6 ist strukturpräferenziell in fast allen Ausprägungen des Motivs durch ein coil- oder ein Strand-Element realisiert. Position 12 und 13 sind dagegen klar durch eine Präferenz zum Helixelement geprägt. Vergleich man diese Positionen mit den dazugehörigen gemittelten Energien ((vgl. hierfür wieder Tabelle 3) wird ersichtlich, dass besonders die Positionen 6 und 12 mit besonders niedrigen (-22,95 bzw. – 21,75) Energien assoziiert sind. Demzufolge besitzt das PS00108-Motiv an diesen Stellen seinen funktionellen Kern r strukturell in den meisten Fällen durch zwei Strand-Stellen, in einigen Fällen durch eine Strand/Helix-Kombination und nahezu niemals als Helix/Helix- beziehungsweise andere Kombinationen realisiert ist.

Das Motiv verhält sich also so, dass auch unterschiedliche Strukturelemente seine Funktion vermitteln können, solange diese energetisch stabil in die Gesamtstruktur eingebettet sind. Die These, dass allein einzelne stabile Strukturelemente eine Funktion vermitteln muss also zugunsten der energetischen Einflüsse der Umgebung erweitert werden.

Eine Übersicht aller positionsweisen Sekundärstrukturpräferenzen aller strukturellen und funktionellen Motive ist im Anhang Teil 3 – Verteilung der Sekundärstrukturelemente in Motiven zu finden.

3.3 Clustering der Motive

Obwohl man, wie im vorangegangen Punkt beschrieben, bereits charakteristische Stellen auf verschiedenen Abstraktionsebenen von Motiven beobachten kann ist die Bildung eines Mittelwertes von Energien eine statistisch gesehen eher schwache Größe. Aus diesem Grund sollen im Folgenden Methoden des hierarchischen Clusterings auf die Motive angewandt werden.

3.3.1 Gruppendifferenzierung durch UPGMA

Mit der im Punkt 2.6.1 beschriebenen Methodik wurden alle Motive geclustert. Der UPGMA-Algorithmus liefert als Ergebnis binäre Bäume, anhand derer die energetischen Abstände einzelner Ausprägungen eines Motivs sichtbar werden.

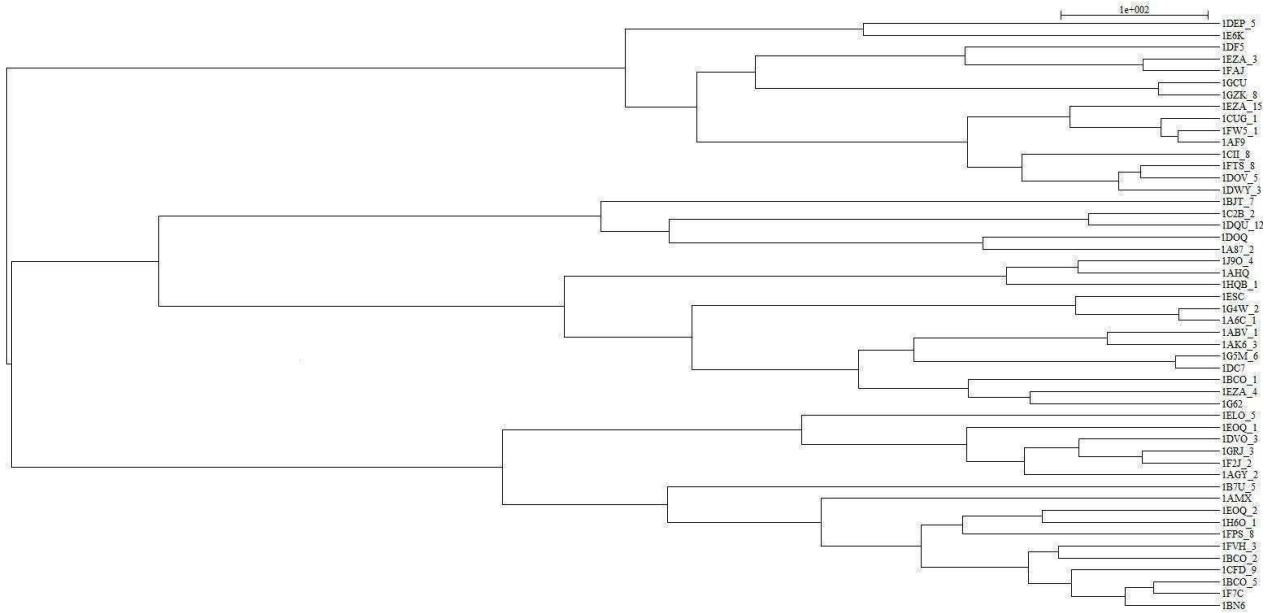


Abbildung 23: Energiebasiertes UPGMA-Clustering für 50 Alphabeta-Motive

Der UPGMA-Baum gibt die Abstände einzelner Ausprägungen eines Motivs wieder und ordnet sie je nach energetischem Abstand zueinander in Gruppen an.

Was im UPGMA-Clustering sichtbar wird ist, dass Motive sich gleicher Art energetisch in Gruppen differenzieren lassen, obwohl sie strukturell klar definiert sind. Die vormals angesprochene Vermutung, dass die sequentielle Realisierung eines Motivs Einfluss auf seine Stabilität im Protein besitzt kann an dieser Stelle also bestätigt werden.

Obige Abbildung zeigt, dass sich die 50 betrachteten Alphabeta-Motive energetisch in mindestens drei Gruppen aufteilen. Zur Validierung des energiebasierten Clusterings lässt sich ebenfalls über den UPGMA-Algorithmus ein Sequenzbasiertes Clustering anfertigen. Als Abstandsmaß dient hier ein normiertes Ähnlichkeitsmaß, welches auf Grundlage der BLOSUM62-Substitutionsmatrix [31] ermittelt wurde.

In einer solchen Substitutionsmatrix sind Werte eingetragen, die angeben, wie wahrscheinlich es ist, dass zwei Aminosäuren in der Sequenz gegeneinander ausgetauscht werden. Diese Wahrscheinlichkeiten dabei sind statistisch-ermittelte Werte.

Sei M_{20x20} eine Substitutionsmatrix (BLOSUM62) und w_i, w_j zwei Sequenzfragmente gleicher Länge n. Weiterhin sei w_{i_k} die k-te Aminosäure im Sequenzfragment w_i. Dann ist $s(w_{i_k}, w_{j_k}) = M(w_{i_k}, w_{j_k})$ der Wert der jeweiligen Aminosäuren in der Substitutionsmatrix. Die Ähnlichkeit Sim zweier Sequenzfragmente ergibt sich dann analog zum euklidischen Abstand als Summe der k-paarweisen Werte in der Matrix:

$$Sim(w_i, w_j) = \sum_{k=1}^{n} M(w_{i_k}, w_{j_k}) \qquad (11)$$

Damit aus diesem Ähnlichkeitsmaß ein Abstand D wird, muss es normiert werden. An dieser Stelle bietet sich die Normierung des erhaltenen Wertes $Sim(w_i, w_j)$ zu den optimalen Werten $Sim(w_i, w_i)$ und $Sim(w_j, w_j)$ an. Eine daraus abgeleitete Distanzfunktion sieht schlussendlich wie folgt aus:

$$D(w_i, w_j) = Sim(w_i, w_i) + Sim(w_j, w_j) - 2\,Sim(w_i, w_j) \qquad (12)$$

Auf Grundlage dieses normierten Ähnlichkeitsmaßes und der BLOSUM62-Matrix lassen sich nun über den UPGMA-Algorithmus binäre Bäume auf Grundlage von Sequenzdaten erstellen. Derartige, aus Sequenzen abgeleitete Bäume sind allerdings nur für strukturelle Motive sinnvoll, da funktionelle Motive auf sequenzieller Ebene zu konserviert sind um signifikante Ergebnisse zu erhalten.

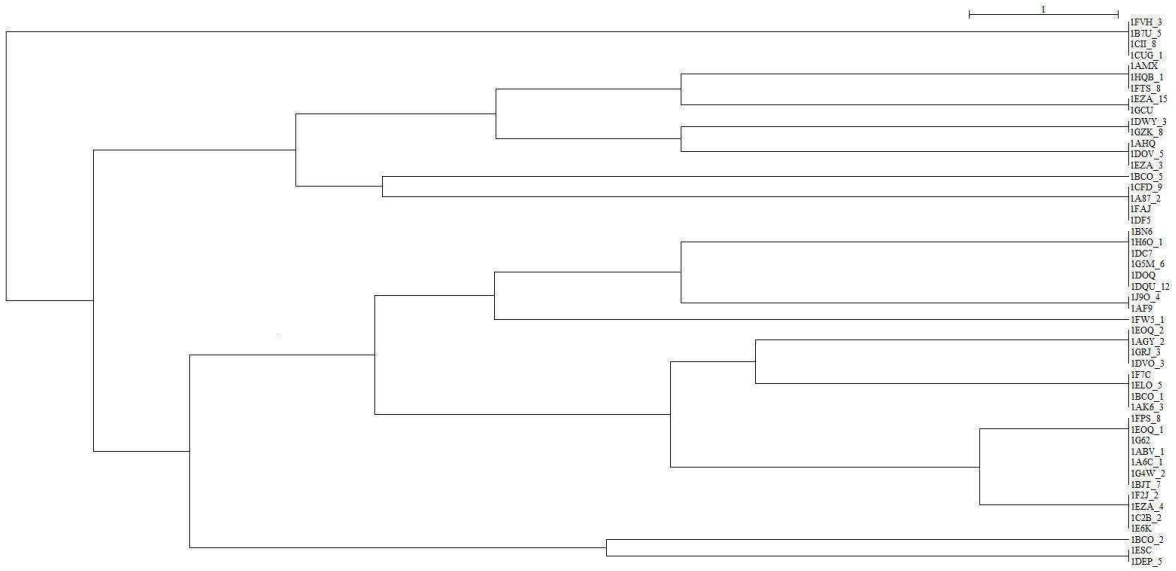


Abbildung 24: Sequenzbasiertes UPGMA-Clustering für 50 Alphabeta-Motive

In diesem UPGMA-Baum sind dieselben 50 Alphabeta-Motive wie in Abbildung 21 geclustert. Als Metrik dient hier allerdings das in (Gleichung 12) formulierte sequenzbasierte Abstandsmaß.

Ein Vergleich der beiden, auf unterschiedlicher Datengrundlage basierenden, Clusterings (Abb. 21 & Abb. 22) zeigt, dass sich die Topologien der Bäume unterscheiden. Auch die Anordnung der Blätter, die die betrachteten Sequenz- beziehungsweise Energiefragmente repräsentieren ist unterschiedlich. Fragmente, die sich im energiebasierten Clustering zu einem Cluster gruppieren sind im sequenzbasierten Clustering oft anders verteilt und umgekehrt.

Analog lassen sich auch Energievektoren funktioneller Motive in einem UPGMA-Clustering darstellen. Dabei ist allerdings zu berücksichtigen, dass nur zwischen Motivausprägungen gleicher Länge sinnvolle Abstände berechnet werden können. Dies ist beispielsweise bei dem Prosite-Motiv PS00007 (Tyrosin-Kinase Phosporylisationsstelle) nicht zwingend der Fall. PS00007 leitet sich aus dem Sequenzmuster [RK]-x(2,3)-[DE]-x(2,3)-Y her, was bedeutet, dass an Position 2 und 4 entweder zwei oder drei beliebige Aminosäuren stehen können. PS00007 ist demnach mindestens 7 und maximal 9 Aminosäuren lang. Ein euklidischer Abstand zwischen zwei ungleich langen Energievektoren eines Motivs kann zwar berechnet werden, allerdings wird durch den Längenunterschied das UPGMA-Clustering verfälscht, da alle nicht vorhandenen Indices in der Distanzmatrix mit dem Wert 0 befüllt werden würden.

Für das Prosite Motiv PS00016 (Zelladhäsionsstelle) wurden 220 seiner Ausprägungen energetisch über den UPGMA-Algorithmus geclustert:

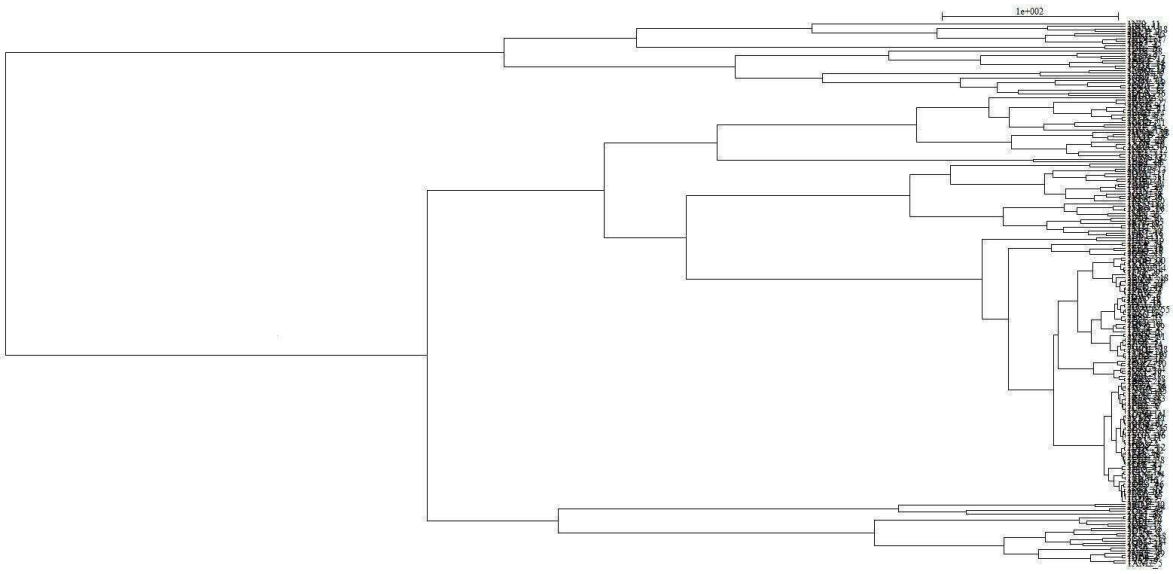


Abbildung 25: Energetisches UPGMA-Clustering für PS00016

Auch im UPGMA-Clustering funktioneller Motive ist erkennbar, dass sich Motive gleicher Funktion energetisch in mindestens drei Gruppen aufteilen. Diese energetische Partitionierung der funktionellen Motive, die sich auf sequentieller Ebene eher homogen verhalten, könnte eine Ursache in der unterschiedlichen strukturellen Realisierung der Motive besitzen.

Ein Problem, dass besonders in obiger Abbildung deutlich wird, ist die zunehmende Komplexität der Bäume mit steigender Zahl seiner Einträge. Das heißt, je mehr Realisierungen eines Motivs untersucht werden, desto komplexer und unübersichtlicher wird das Ergebnis. Dies ist vor allem im Neural-Gas-Clustering angebracht, da in diesem zusätzlich zu den energetischen Abständen der Energievektoren auch deren Energieverläufe visualisiert werden können.

Aus diesem Grund muss ein Verfahren entwickelt werden, dass es erlaubt repräsentative Stichproben aus der Grundgesamtheit des Clusterings zu ziehen. Um das zu erreichen wurde ein intelligentes Monte-Carlo Sampling durchgeführt, in dem alle Cluster voneinander getrennt betrachtet werden. Im Folgenden wird das Sampling am Beispiel der Neural-Gas-Clusterings beschrieben.

Zunächst wird ein zufälliger Datenvektor x_R aus dem vorliegenden Datenraum X gewählt. Dies geschieht dadurch, dass eine Zufallszahl z zwischen Null und Eins gewürfelt und anschließend mit der Gesamtanzahl der Datenvektoren N multipliziert wird. Die Zahl die sich daraus ergibt dient als Index für den Zufallsvektor x_R:

$$x_R := x_i \in X \mid i = z * N \qquad (13)$$

Der Vektor x_R kann prinzipiell in jedem beliebigen Cluster liegen. Anschließend wird die Gesamtanzahl der Datenvektoren des Clusters, in dem sich x_R befindet, ermittelt. Dieser Wert soll C_i genannt werden. Um C_i auf die Gesamtzahl der Datenvektoren im Datenraum zu relativieren wird die Normierungsvariable $A(C_i)$ eingeführt. $A(C_i)$ ergibt sich aus:

$$A(C_i) = \frac{N-|C_i|}{N} \qquad (14)$$

Als nächster Schritt im Sampling wird der Prototyp des Clusters, in dem sich x_R befindet, ermittelt. Von diesem Prototyp P_R ausgehend wird seine energetische Distanz zu allen anderen Prototypen P_i im Datenraum berechnet und aufsummiert:

$$S(P_R) = \sum_{i=0}^{n=|W|} Dist(P_R, P_i) \qquad (15)$$

Die Summe $S(P_R)$ wird anschließend auf die Summe aller Prototypdistanzen $S(P_G)$ normiert (17). Die Matrix DM_P enthält dabei alle Distanzen der Prototypen untereinander.

$$S(P_G) = \sum_{i,j=0}^{n=|W|} DM_P(i,j) \qquad (16)$$

$$S(P_R) = \frac{2}{S(P_G)} \qquad (17)$$

Der letzte Schritt ist die Überprüfung auf einen Schwellwert T:

$$T = 0.2\ A(C_i) * S(P_R) \qquad (18)$$

Die Entscheidungsfunktion E entscheidet schlussendlich über eine Zufallszahl y, ob ein zufällig gewählter Vektor x_R in das Sampling aufgenommen wird, oder nicht:

$$E(x_R) = \begin{cases} 1 & y < T \\ 0 & y > T \end{cases} \qquad (19)$$

Das Ergebnis des Samplings ist eine eigens definierbare Anzahl an repräsentativen Datenvektoren für jedes Cluster. Dabei besitzen Elemente, die sich in großen Clustern befinden eine geringere Wahrscheinlichkeit ausgewählt zu werden, als solche in kleineren, wodurch sicher gestellt wird, dass auch möglicherweise besonders interessante Cluster mit wenigen Elementen im Sampling aufgenommen werden.

3.3.2 Analyse der Clusterenergieverläufe mit Neural Gas

Der Neural Gas Algorithmus eignet sich als numerisch sehr stabiler Vektorquantisieralgorithmus zum Clustern der Energievektoren der Motive besonders in der Hinsicht gut, dass sich die Energieverläufe der geclusterten Motive direkt als Menge von Vektoren dargestellt werden können. Für jedes Motiv wurde eine Iterationszahl von 200000 genutzt. Die Anzahl der Prototypen, die die spätere Anzahl der Cluster repräsentiert, wurde pro Motiv von Hand gewählt.

Im Folgenden sollen die Cluster, die Neural Gas offenlegt miteinander pro Motiv und auch zwischen verschiedenen Motiven verglichen werden.

Für alle 2117 Alphabeta-Motive wurde mit ein Neural-Gas-Clustering erstellt, welches große Unterschiede in den Energieverläufen der Motive offenbart. Initialisiert wurden dabei 15 Prototypen, was am schlussendlich die Anzahl der Cluster auf ebendenselben Wert festlegt.

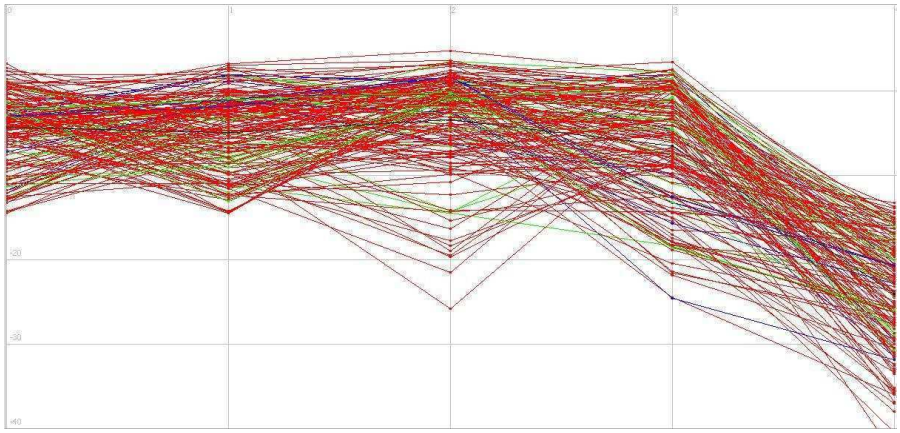


Abbildung 26: Neural-Gas Cluster A für Alphabetamotive

Auf der X-Achse sind im Plot die Positionen im Motiv, auf der Y-Achse die Energieskala mit einem Intervall von [10, -40] aufgetragen. Durch Farben ist das Sekundärstrukturelement codiert, in dem sich der jeweilige Energievektor befindet.

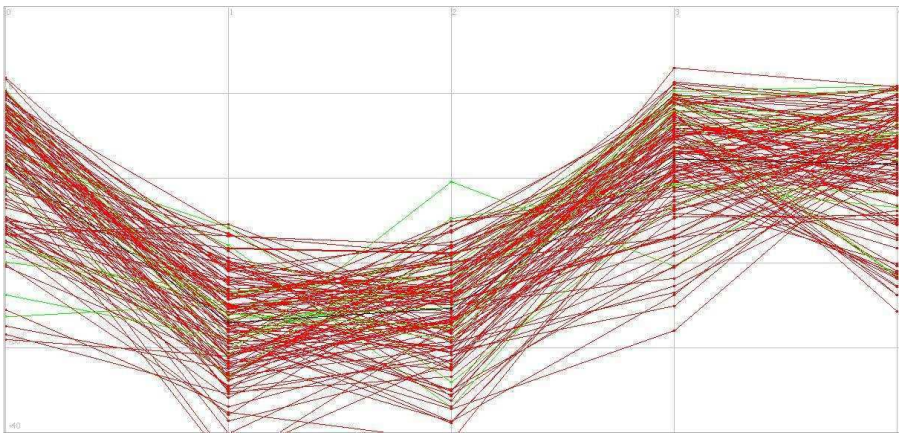


Abbildung 27: Neural-Gas Cluster B für Alphabeta.Motive

Analog zu Abbildung 26 sind hier die Energieverläufe eines weiteren Clusters der Alphabeta.Motive aufgetragen.

Beide Cluster A und B unterscheiden sich hinsichtlich ihrer Energieverläufe stark. Während die dem Cluster A zugeordneten Vektoren einen auf den ersten vier Positionen konstanten und zum Ende hin abfallenden Verlauf haben, verhalten sich die des Clusters B auf Position 2 und 3 energetisch am niedrigsten.

Die sichtbaren Energieminima können an beinahe jeder Stelle im Motiv auftreten, was eine Besonderheit dieser Positionen impliziert. Dieses Verhalten kann Ursachen in der sequentiellen Realisierung der jeweiligen Motive besitzen, oder aber an Umgebungseigenschaften geknüpft sein, also an strukturelle Besonderheiten außerhalt des eigentlichen Motivs.

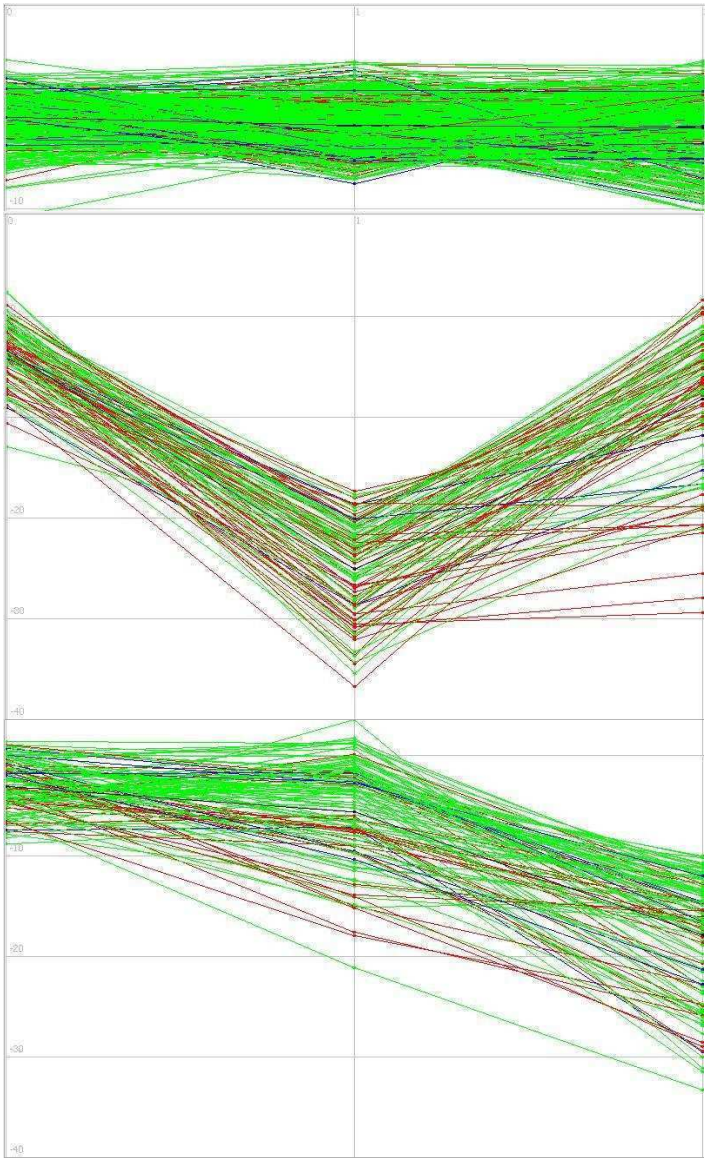


Abbildung 28: Energieverläufe der Neural-Gas Cluster für Asx-turns

Die Energieverläufe von 907 Asx-turns lassen zu, dass sich das Motiv energetisch in drei klare Gruppen einteilen lässt.

Während obige Gruppe zumeist in coil-Bereichen (grün) lokalisiert ist, treten die beiden restlichen Gruppen zu einem gewissen Anteil auch in Helix- (rot) und Strand-Bereichen (blau) auf, was auch deren insgesamt niedrigeren Energieverläufe erklärt.

Eine Besonderheit, die sich ergibt, wenn die Energieverläufe der Neural-Gas Cluster betrachtet werden ist, dass sich zwar jedes Motiv in charakteristische Energieverläufe unterscheiden lässt, die Menge dieser Unterscheidungen allerdings Überlappungen zwischen Motiven erzeugt.

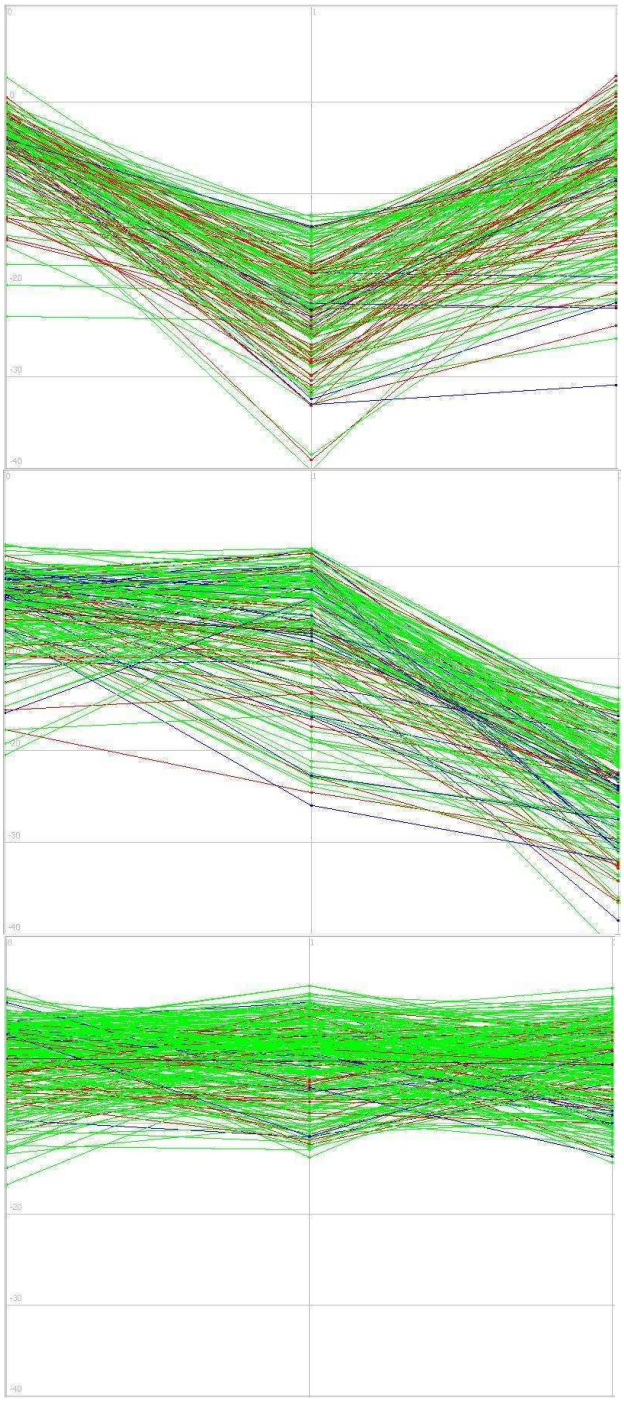


Abbildung 29: Energieverläufe der Neural-Gas Cluster für Nest-Motive

Die Energievektoren von 694 Nest-Motive gleichen denen von Asx-turns an allen Positionen

Das Problem, welches an dieser Stelle der Ausführungen offensichtlich wird, ist die nach der Frage der Detektierbarkeit der Motive. Obgleich es sich bei Asx-turns und Nests um strukturell klar voneinander abzugrenzende Motive handelt, sind ihre Energieverläufe nahezu gleich. Dieses Phänomen lässt sich beispielsweise ebenfalls bei Asx- und Alphabeta- Motiven und weiteren beobachten. Weiterhin ist festzuhalten, dass charakteristische Energieverläufe kleinerer Motive in die größerer Motive hineinzupassen scheinen (vgl. dazu Abbildung 26 und 28 bzw. 29). Die naheliegende Vermutung, dass kleine strukturelle Motive größere aufzubauen vermögen kann allerdings nicht bestätigt werden. Selbige Differenzierung der Strukturmotive in ihren Energieverläufen lässt sich auch bei funktionellen Motiven beobachten, die wiederrum ähnliche Energieverläufe untereinander, als auch zu den Strukturmotiven aufweisen können.

4 Interpretation der Clusterings

Jedes Clustering, dass durchgeführt wurde, sei dies nun mit dem UPGMA- oder dem Neural-Gas-Algorithmus erstellt, zeigt die Abstände verschiedener Energievektoren von Motiven einer Art. Dies bedeutet, wenn sich Cluster ausbilden, muss es signifikante Unterschiede in den Energieverläufen von Motiven gleicher Art geben.

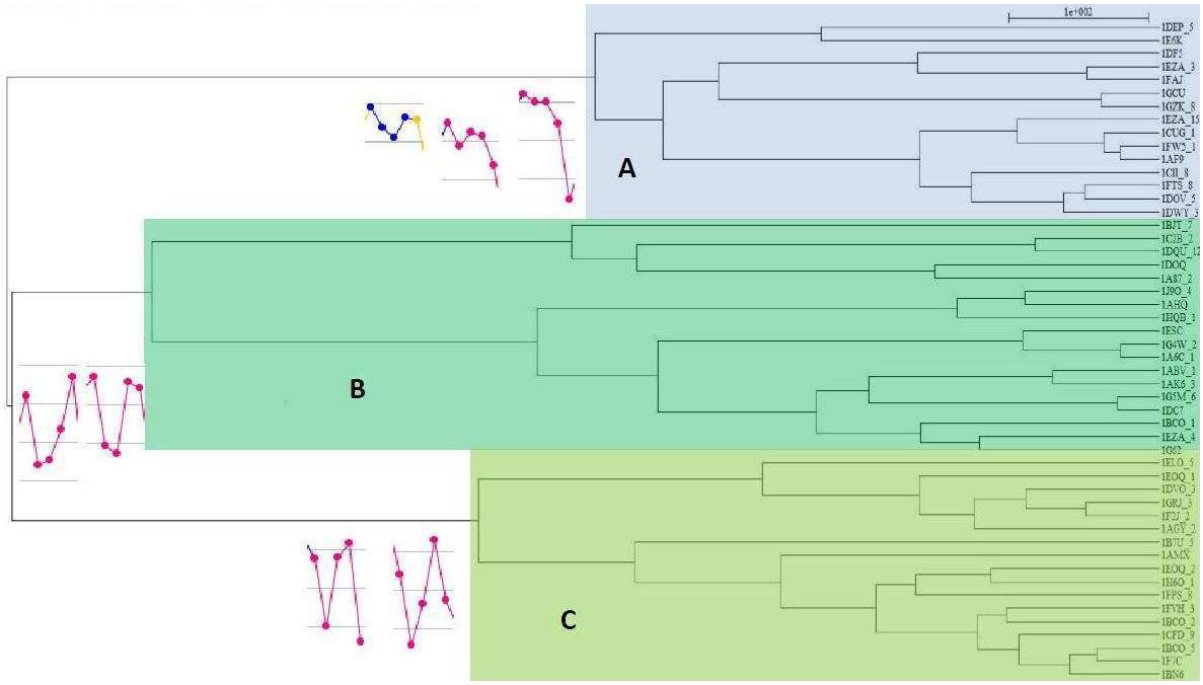


Abbildung 30: Energieverläufe einzelner Alphabetamotiv-Cluster

Ein UPGMA-Clustering für 50 gesampelte Alphabeta-Motive. Farbig markiert sind hier zunächst willkürlich festgelegte Clustergrenzen. Nebenstehend sind die Energieverläufe einiger Einträge der einzelnen Cluster abgebildet.

Untersucht man einzelne Vertreter der drei Cluster A, B und C lassen sich deren Energieverläufe visualisieren und Unterschiede feststellen. Alphabetamotive sind Motive mit einer Länge von fünf Aminosäuren, was impliziert, dass auch seine Energieverläufe durch fünf Energiewerte gebildet werden. Aus dem Energieprofil der Gesamtstruktur können entsprechende Fragmente ausgeschnitten werden.

Das obere Cluster A verhält sich so, dass die Energieverläufe seiner Einträge gegen Ende des Motivs abfallend sind. Cluster B hingegen besitzt sein Energieminimum auf der zweiten und dritten Position im Motiv. Das letzte Cluster C ist zu B ähnlicher als zu A, was sowohl in der Topologie des Baumes, als auch in den Energieverläufen deutlich wird. Es besitzt ähnlich wie Cluster B ein Energieminimum an zweiter Position, geht dann allerdings auf der nächsten Position direkt in ein Maximum über, um schlussendlich wieder abzufallen.

Diese Logik funktioniert ebenso andersherum. Betrachtet man beispielsweise die Energieverläufe 50 gesampelter Asx-turns, so lassen diese sich in drei, klar voneinander abgrenzbare Grundtypen unterscheiden:

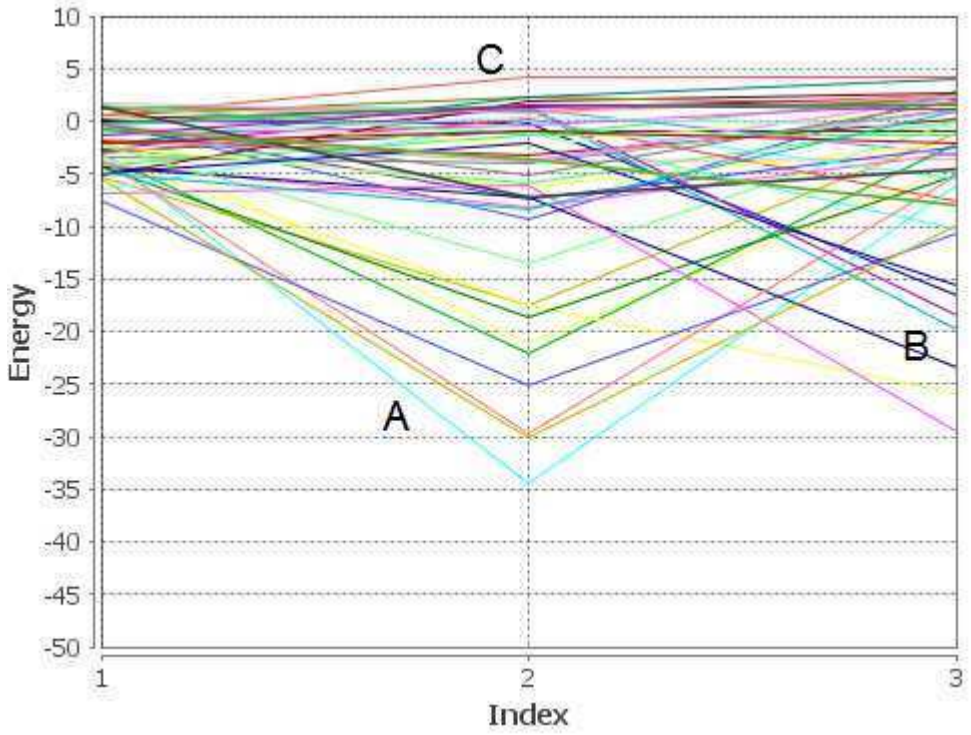


Abbildung 31: Energieverläufe 50 gesampelter Asx-turns

Trägt man die Energieverläufe mehrerer Ausprägungen eines Motivs aufeinander ab, lassen sich auch hier bereits klare Grupperungen erkennen (vgl. Abbildung 28).

Typ A ließe sich als alternierend in seinem Energieverlauf beschreiben, während Typ B auf den ersten beiden Positionen leicht anwächst, um dann an letzter Stelle stark abzufallen. Typ C verhält sich auf seiner gesamten Länge nahezu konstant.

Im UPGMA-Clustering exakt derselben Motive wird diese Unterscheidung wiederum deutlich:

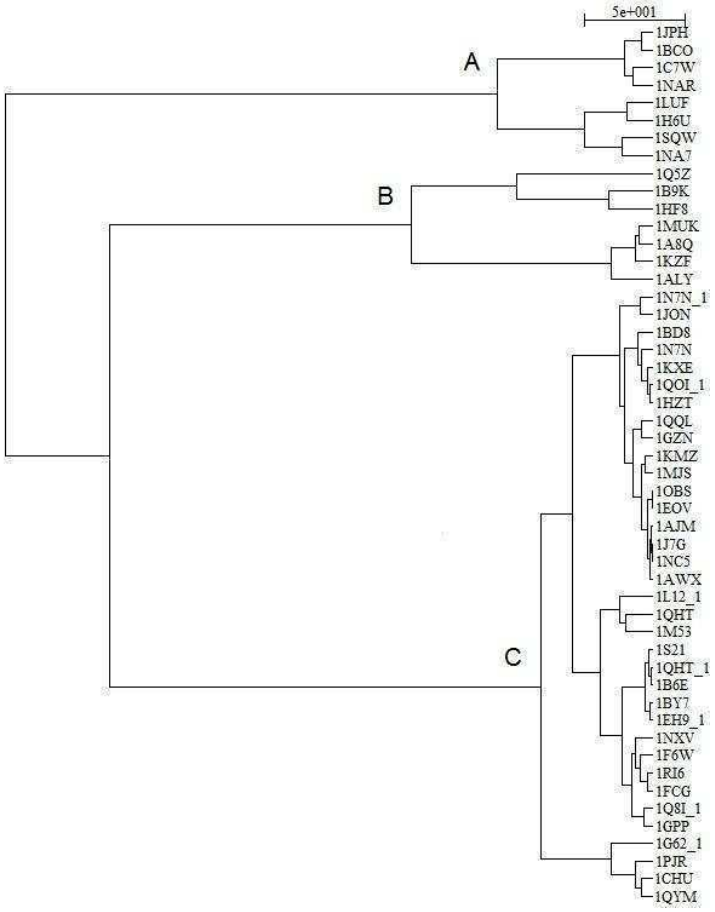


Abbildung 32: UPGMA-Clustering 50 gesampelter Asx-turns

Das UPGMA-Clustering von 50 gesampelten Asx-turns zeigt starke Parallelen zu den Energieverläufen derselben Motive. Das größte Cluster C repräsentiert alle Motive mit einem weitgehend konstanten Energieverlauf.

Gleicht man die ID's der Motive im UPGMA-Baum mit ihren Energieverläufen ab, so zeigen sich die eindeutigen Parallelen zu Abbildung 27. Die UPGMA-Clusterings können somit als valide angesehen werden, da die Energieverläufe der Motive sich in dieselben Gruppen einteilen lassen, wie es die Cluster im Clustering tun. Die soeben gezeigten Grafiken und Clusterings sollen für eine große Menge dieser Art von Daten sprechen, von denen der Großteil den ausgeführten Gesetzmäßigkeiten folgt.

Werden die Clusterings für funktionelle Motive betrachtet, so müssen diese ein wenig anders interpretiert werden. Zwar zeichnen sich auch in ihnen die charakteristischen Energieverläufe der einzelnen Cluster ab, jedoch sind deren Unterschiede wiederum weniger in der sequentiellen, sondern vielmehr in der strukturellen Realisierung der Motive zu suchen. Dies ist dann der Fall, wenn eine Aminosäure in einem funktionellen Motiv über das Sequenzpattern eindeutig determiniert ist. Beispielsweise im Motiv PS00016, welches durch das Sequenzpattern R-G-D definiert ist, ist zwar die sequentielle Realisierung in jeder Ausprägung gleich, jedoch nicht seine strukturelle. Es verhält sich vielmehr so, dass jede Position viele mögliche Winkelkonformationen einnehmen kann, was wiederum die energetische Stabilität des Motivs beeinflusst. Die funktionellen Motive stehen in dieser Hinsicht also im exakten Gegensatz zu Strukturellen.

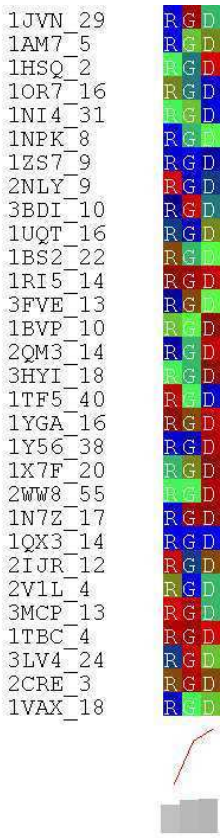


Abbildung 33: Energetisch-/sequentielles Verhalten gesampelter PS00016-Motive

Aufgetragen sind hier 30 gesampelte Ausprägungen des PS00016-Motivs. Durch Farben codiert ist dabei die Größe der Energie der jeweiligen Position, wobei rot einen hochenergetischen, blau einen niederenergetischen und grün die Energiebereiche dazwischen darstellt.

In Abbildung 29 ist ersichtlich, dass jede Position, obgleich sie sich sequentiell exakt gleichen, energetisch beinahe jeden Zustand einnehmen kann. Dies kann nur eine Ursache in der jeweiligen strukturellen Ausrichtung einer Aminosäure in der Gesamtstruktur haben.

Weiterhin ist ersichtlich, dass es Realisierungen von PS00016 gibt, die entweder energetisch insgesamt im hohen Energiebereich, (siehe in Abb. 29 ID: 1TBC_4) oder in niedrigeren Energiebereichen (siehe 1QX3_14) liegen. Andere wiederrum erstrecken sich an allen Positionen über ein größeres Energiespektrum (siehe 1QM3_14). Die Energieverläufe unterscheiden sich folglich auch hier zum Teil enorm, was sich wiederrum im UPGMA-Clustering zeigt:

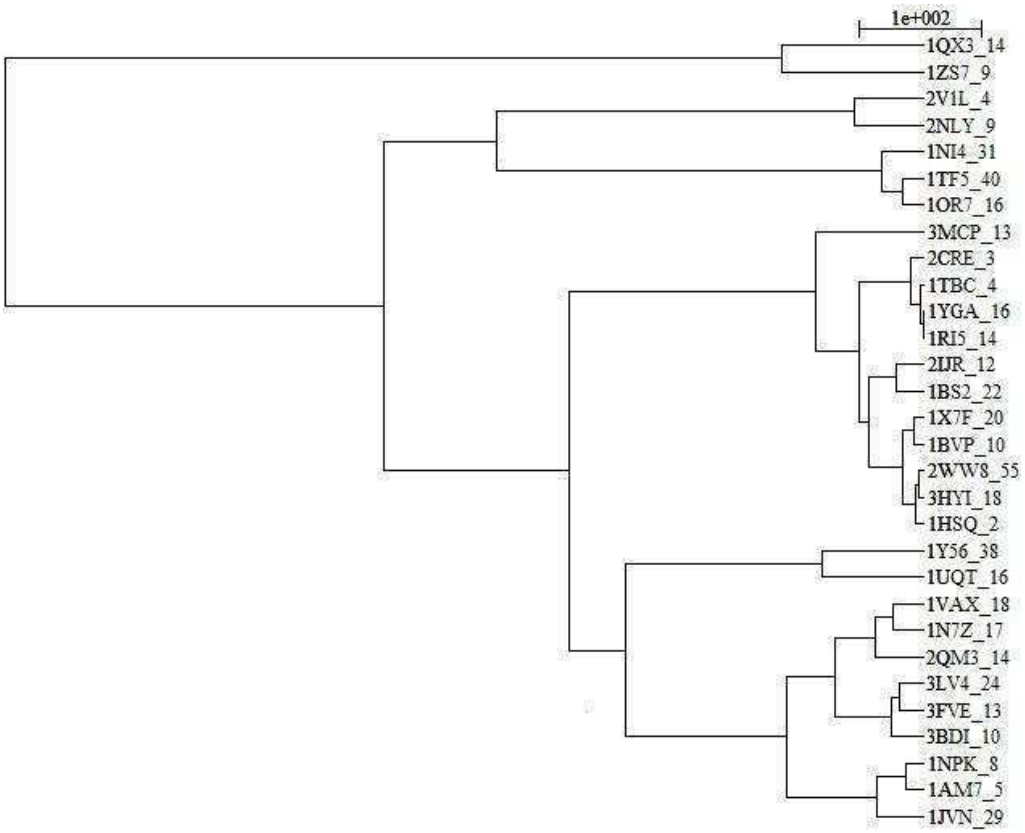


Abbildung 34: UPGMA-Clustering für 30 gesampelte PS00016-Motive

Die Aufteilung der Elemente im UPGMA-Clustering lässt sich direkt mit den energetischen Charakteristika derselben Elemente in Abbildung 29 in Verbindung bringen.

Die Motiv-Ausprägung 1QX3_14 bildet zusammen mit 1ZS7_9 eine Außengruppe im Clustering, da beide über alle Positionen energetisch niedrig liegen (siehe Abbildung 29). Andere höher-energetische Ausprägungen wie beispielsweise 1TBC_4 liegen distanziell im Clustering sehr weit entfernt von diesen beiden.

Die Frage, die sich nun stellt, ist die nach der Ursache der energetischen Differenzierung der strukturellen und funktionellen Motive. Wie ist es möglich, dass Strukturmotive, die sich strukturell nur in engen Bereichen bewegen und funktionelle Motive, die selbiges Verhalten auf sequentieller Ebene aufweisen, sich im energiebasierten Clustering teilweise weit voneinander entfernen. Weiterhin stellt sich die Frage, wie funktionelle Motive ihre Funktion ausführen können, wenn sie sich energetisch weit differenzieren lassen. Ein sehr passendes Beispiel hierfür ist das eben betrachtete PS00016-Motiv. Sequentiell, sowie funktionell ist es klar definiert. Energetisch und somit auch strukturell scheint es dagegen allerdings Unterschiede zu geben. Die Funktionsweise des Motivs muss also auf mehreren Abstraktionsebenen determiniert sein. Offenbar gibt es strukturelle Variationen innerhalb eines funktionellen Motivs, die es dennoch zulassen, dass das Motiv seine biologische Funktion erfüllen kann.

Im folgenden Punkt soll nun der Frage nachgegangen werden, was der Grund für die mannigfaltige Differenzierung der Motive auf energetischer Ebene ist.

5 Strukturbiologische Clusteranalyse

Auf struktureller Ebene sind Strukturmotive selbsterklärend als hochkonserviert zu bewerten. Die Unterschiede in den Energieverläufen unterschiedlicher Cluster müssen demnach eine Folge ihrer sequentiellen Variabilität sein. Die physikochemischen Eigenschaften eines strukturellen Motivs variieren mit seiner sequentiellen Ausprägung. Während einige wie Asx- oder St-Motive an manchen Positionen sequentiell determiniert sind, ist dies beim Großteil der übrigen Motive nicht der Fall.

Die im Punkt 3.3.2 angesprochenen energetischen Überlappungen zwischen Motiven lassen den Schluss zu, dass kleine strukturelle Motive größere funktionelle Motive aufbauen können oder zumindest mit diesen assoziiert sein können. Die energetischen Analysen legen offen, dass es durchaus möglich ist, dass ein strukturelles Motiv ein funktionelles aufbaut. So gesehen können kurze strukturelle Motive auch als Bauelemente für größere funktionelle Motive betrachtet werden. Dies wäre beispielsweise beim PS00107-Motiv der Fall. Seine – wahrscheinlich – funktionell bedeutsame Affinität zu Strand-Bereichen an seinen Enden (vgl. hierzu Anlagen, Teil 3 – Verteilung der Sekundärstrukturelemente in Motiven und Tabelle 4 - Mittelwerte der positionsweisen Energien für funktionelle Motive) impliziert, dass an den entsprechenden Stellen auch faltblatttypische Strukturmotive wie Betabulges oder Betaturns zu finden sind.

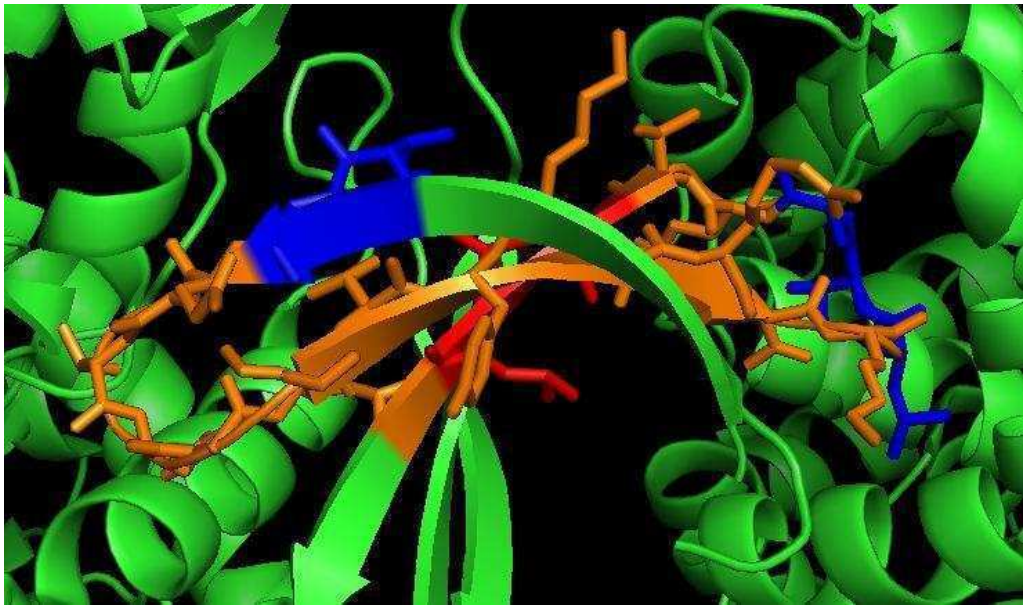


Abbildung 35: Flankierende Betamotive an einem PS00108-Motiv

Diese Ausprägung des PS00107-Motivs in 2C5N wird von einem Betabulge (links) und einem Betaturn (rechts) flankiert und stabilisiert.

In der Struktur 2C5N wird ein PS00107-Motiv (orange dargestellt) von einem Betabulge (blau) und einem Betaturn (ebenfalls blau) flankiert. Die Betaturns betten das funktionelle Motiv dabei in die Struktur ein. Das aufgrund sehr niedriger Energien (-28.05, -21.25, -30.64) wahrscheinliste aktive Zentrum des Motivs (rot) wird also durch die Beta-Motive in der Gesamtstruktur gehalten.

Im Falle des PS00108-Motivs verhält es sich so, dass zwei wesentliche Cluster unterschieden werden können:

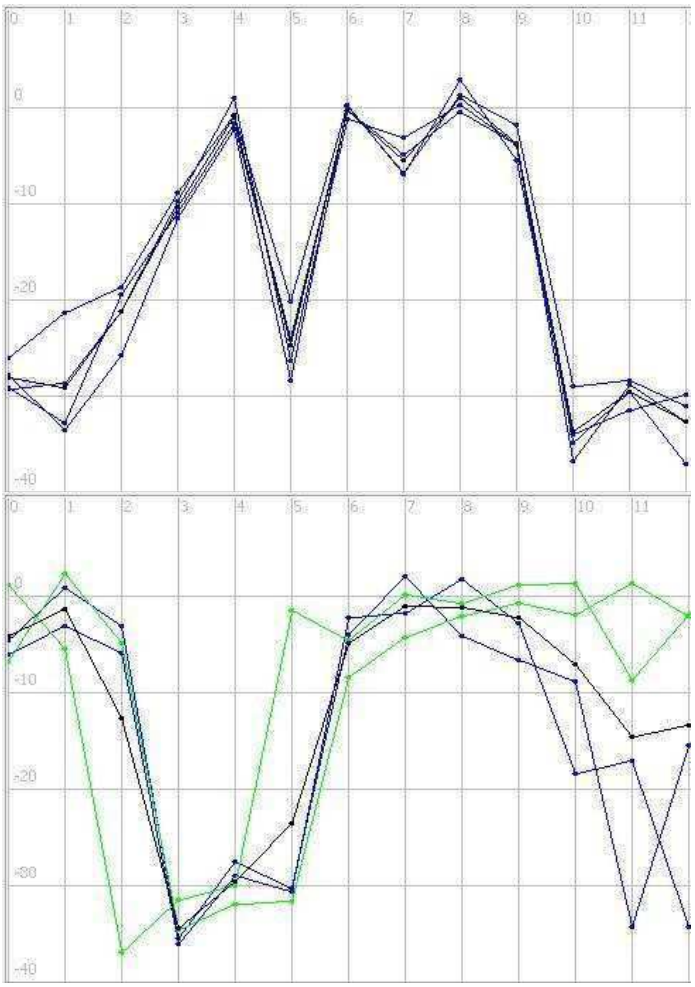


Abbildung 36: Neural-Gas Cluster von PS00108

Das PS00108-Motiv kann in zwei wesentliche Cluster unterschieden werden. Obiges Cluster A besitzt ein Energiemaximum dort, wo unteres Cluster B durch zwei Energieminima charakterisiert ist.

Die Ursache hierfür ist, dass Strukturen, die ins Cluster A fallen (beispielsweise 2QVS oder 1A06) an Position 4 mit einer kurzen Helix assoziiert sind, während solche, die dem Cluster B zugehörig sind (bspw. 1LPU), zusätzlich zu dieser kurzen Helix noch mit einem Faltblatt belegt sind. Der Unterschied zwischen energetischen Clustern von funktionellen Motiven ist also in deren struktureller Realisierung zu finden.

Exemplarisch für die strukturellen Motive soll eine Clusteranalyse für alle Alphabeta-Motive anhand folgenden Clusterings dienen:

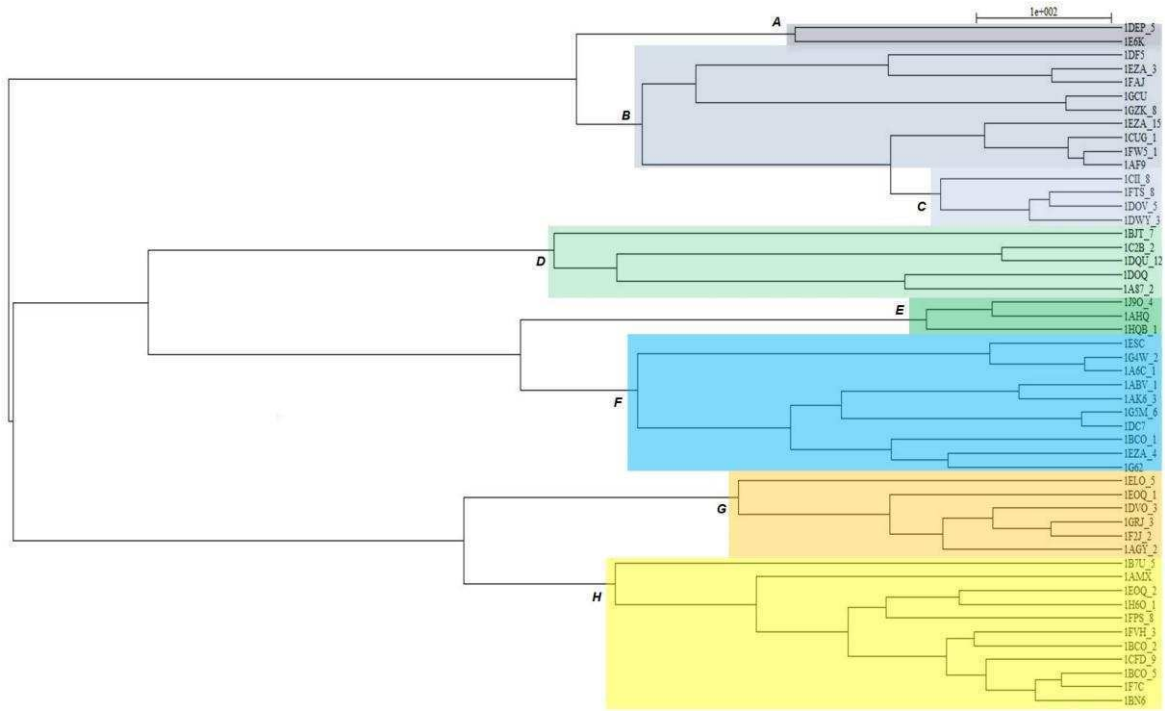


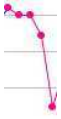
Abbildung 37: Festlegung der Cluster für Alphabeta-Motive

Für 50 gesampelte Alphabeta-Motive wurden Cluster ausgewählt, die auf ihre sequentiellen, sowie sekundärstrukturellen Eigenschaften hin überprüft werden sollen.

Die folgenden Tabellen sollen für eine Reihe im obenstehenden UPGMA-Clustering stehenden Einträge sequentielle sowie sekundärstrukturelle Charakteristika der jeweiligen Position im Motiv mit den entsprechenden Energien in Verbindung bringen.

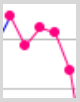
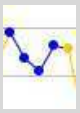
Cluster A

Tabelle 10: Clusteranalyse für Alphabeta-Motive in Cluster A

ID	Energieverlauf	Energievektor	Sequenz	Sekundärstruktur-elemente
1DEP_5		-12.568211114691604 -3.3009814692442228 0.8236813890123544 -9.160790637082783 -16.837108502426943	FRKAF	HHHHH
1E6K		0.10458092098030303 0.05556137643051634 -5.914274645135152 -25.428003179602275 -15.393490433112243	KKENI	HHHHH

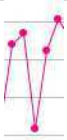
Cluster B

Tabelle 11: Clusteranalyse für Alphabeta-Motive in Cluster B

ID	Energieverlauf	Energievektor	Sequenz	Sekundärstruktur-elemente
1DF5		-12.506225929888705 -13.26555182285579 -0.16599131441095294 -11.432709819377164 -0.44412385383566866	WMEWD	HHHHH
1FAJ		0.6750431787934692 -21.237307460505217 -6.6512260172293365 1.9994976455849467 0.5544893888502991	DLPED	ccccc
1GZK_8		4.682125479538182 -1.6175214065126293 2.4397188903416938 1.4414656986209826 -6.389459886551595	DPKQR	HHHHH
1CUG_1		-2.4102362232015495 -2.4452698903113035 -15.294787025349677 -3.038817729532746 -1.0107796547140433	PDARG	HHHHc
1AF9		-1.0190689598729106 -6.4636042742123525 -9.035086441664514 -3.7985675263128553 -4.4535611970218625	DRLSS	ccccS

Cluster C

Tabelle 12: Clusteranalyse für Alphabeta-Motive in Cluster C

ID	Energieverlauf	Energievektor	Sequenz	Sekundärstruktur-elemente
1CII_8		-9.183456443603065 -4.985630950966934 -24.336875078130568 -17.134278133543052 -5.194654567325015	ARLLG	HHHHH
1FTS_8		-6.052720373664482 -3.1333158986898866 -28.466355219720878 -7.7556962465624455 0.8089422084783708	TNLTE	HHHHH
1DOV_5		-1.4170387294213964 -0.4178746658043937 -24.61770768575642 0.12268723694330985 -0.29864150042259885	PEVDK	HHHHH
1DWY_3		-10.061238317002198 -8.863763936540074 -23.322457412655428 -2.2013853187248302 -2.350391636654706	TQYQR	HHHHH

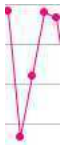
Cluster D

Tabelle 13:Clusteranalyse für Alphabeta-Motive in Cluster D

ID	Energieverlauf	Energievektor	Sequenz	Sekundärstruktur-elemente
1BJT_7		-17.396005827012925 -29.877303554094148 - 43.023188690822835 -34.01569530725548 -29.827484247643277	PMILV	cHHHH
1DQU_12		-20.495878445815503 -20.00230052670269 -41.287406838964976 -11.230199337230655 -18.06212347786537	VAIRA	HHHHH
1DOQ		-8.724690238055981 -19.487256226143497 -42.736628114997764 -28.959829779831026 -11.969301584368218	RALLA	HHHHH
1A87_2		-14.998781276835492 -17.57539663361892 -29.525012784960595 -29.805220487666364 -1.8272302806916456	SSFID	HHHHc

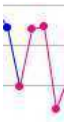
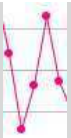
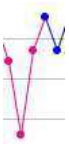
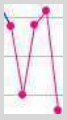
Cluster E

Tabelle 14: Clusteranalyse für Alphabeta-Motive in Cluster E

ID	Energieverlauf	Energievektor	Sequenz	Sekundärstruk-turelemente
1J9O_4		-2.790896755519101 -20.838063040625222 -22.954919937509278 -4.24680140474252 -5.894211565428802	DVVRS	HHHHH
1AHQ		-1.7575343118021949 -33.417154095099484 -18.23736830452267 -2.1667587912705013 -3.4063220568962076	DCVQK	HHHHH
1HQB_1		-8.008007251003024 -26.124789523220247 -24.688394079364894 -16.950570330296273 -3.1296378937442775	QLLLE	HHHHH

Cluster F

Tabelle 15: Clusteranalyse für Alphabeta-Motive in Cluster F

ID	Energieverlauf	Energievektor	Sequenz	Sekundärstruk-turelemente
1ESC		1.1398020116199423 -31.402575960777966 -10.214986301118634 -0.9920319782255186 -24.55336002353831	KCGEF	cHHHH
1A6C_1		-5.772167553455402 -20.34986802996192 -6.1141097547431205 -5.396987970332743 -26.28591536075929	TLRQV	cHHHH
1AK6_3		-16.243432089353046 -34.56468954627218 -23.72446602735285 -7.0169137022322055 -22.855014591029008	HFVGM	HHHHH
1DC7		-15.837656070851114 -34.09884737040702 -13.247208870253129 -4.776911648059725 -13.036276600770588	ALAGA	HHHcc
1G62		-12.534047744285791 -30.158679510194133 -12.050009633371946 -8.368447923198692 -34.2188769968013	ALGNV	HHHHH

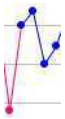
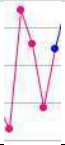
Cluster G

Tabelle 16: Clusteranalyse für Alphabeta-Motive in Cluster G

ID	Energieverlauf	Energievektor	Sequenz	Sekundärstruk-turelemente
1ELO_5		-33.03217637657607 0.5341416168291404 1.8029103047858717 -17.819247705602983 -25.551801668011713	IEEAM	HHHHH
1DVO_3		-20.5667339557507 -22.610962485245686 -0.3258387413050814 -8.433354710314205 -31.244985117568014	AVNTL	HHHHH
1F2J_2		-26.436672505338862 -12.875520208175091 -5.616997336934781 -16.26864892156797 -30.499162187223767	LARVI	HHHHH
1AGY_2		-18.72865452950152 -9.282543961541538 -0.5146907972896265 -20.327794265954505 -34.327468849969044	APEFL	HHHHH

Cluster H

Tabelle 17: Clusteranalyse für Alphabeta-Motive in Cluster H

ID	Energieverlauf	Energievektor	Sequenz	Sekundärstruk-turelemente
1B7U_5		-34.3437034288803 -4.278238621632142 -4.410296333328775 -8.277317404856696 -4.130162743438179	FNQTG	HHccc
1EOQ_2		-22.847659756371396 -2.455719976976428 -25.14347875362645 -26.54784684451816 -14.40081125833613	IKYVL	HHHHH
1FPS_8		-24.61762542295025 1.5393953813416594 -12.624049772900591 -26.65550245788748 -23.47667231567666	YKAIV	HHHHH
1BCO_2		-32.264505462621585 -10.101623525004353 -6.340258180358494 -20.16409085588243 -15.393245415683513	LAGAY	Hcccc
1BCO_5		-28.79169822862153 -11.036889168264635 -17.727073679983466 -28.888154273379044 -8.823903745144914	VAMFN	HHHHH
1BN6		-37.0927629826035 -5.624496766028344 -14.52261225034371 -31.39238582786387 -15.931216634349639	IPHVA	HHHHc

Das Alphabeta-Motiv hält sich als Helix-typisches Strukturmotiv erwartungsgemäß in der Masse seiner Ausprägungen auch in helikalen Strukturbereichen (H) auf. Seltene Positionen, die in coil- (c) oder Strandbereichen (S) lokalisiert sind, sind tendenziell auch mit einer höheren Energie an dieser Stelle realisiert. Die Cluster A und B lassen sich – im Vergleich zu den anderen Clustern – als hochenergetisch und ohne Position mit einem klaren Energieminimum beschreiben. So gesehen können diese beiden Cluster zu einem, sich variabel verhaltenden, Cluster zusammengefasst werden.

Die Cluster C, D und E lassen sich ebenfalls zu einem Cluster zusammenfassen, dass an mittlerer Position sein Energieminimum besitzt. Im Falle des E-Clusters erstreckt sich dieses Minimum allerdings über zwei Positionen, was seine besondere Stellung im UPGMA-Baum erklärt.

Warum das Cluster C im UPGMA-Baum dem B-Cluster untergeordnet ist, obwohl es vektoriell dem D-Cluster näher ist, lässt sich aus dieser Betrachtung nicht sagen. Es lässt sich allerdings feststellen, warum die Cluster C, D und E ihr Energieminimum an den mittleren Positionen besitzen. Sie bestehen an diesen Stellen stets aus Aminosäuren mit hydrophobem Charakter wie Leucin, Isoleucin, Valin oder Phenylalanin, was suggeriert, dass sich diese Positionen im Proteininneren befinden und dort helikale Strukturen ausbilden.

Das F-Cluster besitzt sein Energieminimum an zweiter Position und ist in vier von fünf Ausprägungen an seinen Enden mit coil-Strukturen assoziiert. Es scheint also so, dass die Alphabeta-Motive dieses Cluster den Anfang beziehungsweise das Ende einer Helix bilden. Da diese coil-Positionen ebenfalls mit hydrophoben Aminosäuren wie Glycin, Alanin, Lysin oder Threonin belegt sind, lässt sich schlussfolgern, dass es sich auch hier um im Inneren des Proteins lokalisierte Strukturelemente handelt.

Distanziell im Baum weit von den restlichen Clustern entfernt, befinden sich das G- und das H-Cluster. Alphabeta-Motive, die dem G-Cluster zugehörig sind zeichnen sich dadurch aus, dass sie an erster beziehungsweise letzter Position mit sehr niedrigen Energien belegt sind. Es lässt sich daher vermuten, dass diese Stellen möglicherweise auch funktionell bedeutsam sind. In der Struktur von 1ELO lässt sich diese Vermutung belegen:

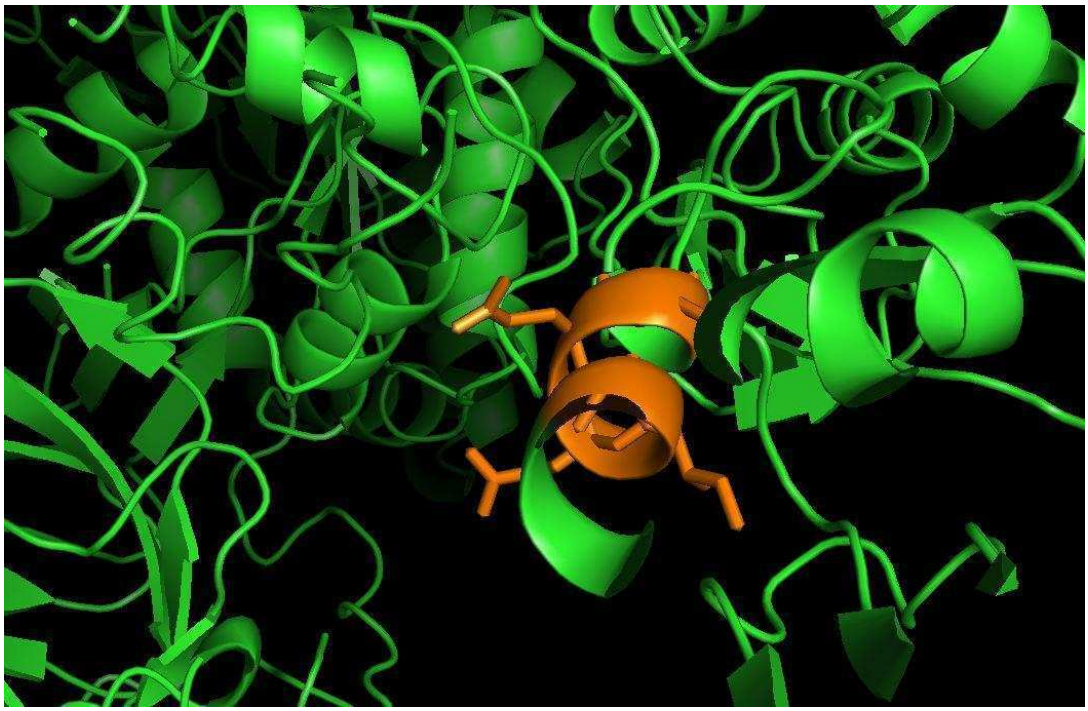


Abbildung 38: Alphabeta-Motiv mit PS00008 assoziiert

In der Struktur 1ELO wird das funktionelle Motiv PS00008 durch ein Alphabeta-Motiv des G-Clusters gebildet.

Ähnlich wie PS00108 verhält sich das PS00008-Motiv strukturell ambivalent (vgl. hierzu Anlagen, Teil 3 – Verteilung der Sekundärstrukturelemente in Motiven). Ist also eine Ausprägung des PS00008-Motivs durch eine helikale Sekundärstruktur realisiert, lässt sich daraus schließen, dass diese von einem Alphabeta-Motiv des G-Clusters gebildet wird.

Energetisch nah am G-Cluster befindlich, lässt sich der ausgeführte Struktur-Energie-Funktions-Zusammenhang allerdings nicht auf das H-Cluster anwenden. Obwohl es analog zum G-Cluster an erster Position ebenfalls mit sehr niedrigen Energien belegt ist, ist es funktionell nicht bedeutsam. Aufgrund seiner Affinität zu coil-Bereichen an seinen Enden ist es wahrscheinlich eher als Brückenelement zwischen coil-Strukturen und besonders stabilen Helices zu interpretieren.

6 Diskussion der Detektierbarkeit von Motiven

Im nun folgenden Kapitel sollen die Erkenntnisse der vorigen Kapitel kritisch bewertet und in einen größeren Kontext gestellt werden. Es soll dabei vor allem auf die Frage der Detektierbarkeit der strukturellen und funktionellen Motive eingegangen werden.

Die Analysen der Clusterings und ausgewählter Strukturen haben gezeigt, dass es Zusammenhänge zwischen charakteristischen Energieverläufen, sekundärstrukturellen Präferenzen und Funktionen zwischen Motiven gleicher und verschiedener Art gibt. Ein Problem, welches sich offenbart, wenn Energieverläufe verglichen werden ist, dass der energetischen Überlappungen zwischen Motiven. So würde ein Algorithmus, der ein Energieprofil beispielsweise auf das Vorkommen von Asx-turns überprüfen soll viele falsch-positive Ergebnisse liefern, da die charakteristischen Energieverläufe von Asx-turns und beispielsweise Nests nahezu gleich sind. Die Anzahl der falsch-positiven Ergebnisse steigt weiter, wenn bedacht wird, dass die Energievektoren kleiner struktureller Motive durchaus Teilvektoren größerer Motive sein können, auch wenn sie strukturell in keiner Verbindung stehen.

Weiterhin spielt die Größe eines Motivs eine Rolle dabei, wie gut es detektiert werden kann. Ein Betabulge mit einer Länge von lediglich 2 Aminosäuren ist beispielsweise viel zu kurz um effizient und richtig detektiert zu werden. Seine Energieverläufe sind für diese Aufgabe zu unspezifisch.

Ein Vorteil den die funktionellen Motive gegenüber den strukturellen besitzen ist in dieser Hinsicht ihre Länge. Der im Punkt 5 angesprochene Zusammenhang von Strukturmotiven und größeren funktionellen Motiven lässt die Überlegung zu, dass funktionelle Motive anhand von charakteristischen Strukturmotiven detektiert werden können.

Den strukturellen Kern, eines in helikalen Bereichen lokalisiertem, PS00008-Motiv stellt ein Alphabeta-Motiv dar. Da das Alphabeta-Motiv jedoch energetisch hochvariabel ist, muss die Auswahl in der Art eingeschränkt werden, dass als Erkennungskriterium nur die Energieverläufe von Alphabeta-Motiven des G-Clusters dienen (vgl. hierzu Tabelle 18: Clusteranalyse für Alphabeta-Motive in Cluster G). Aus deren charakteristischen Energieverläufen lassen sich nun energetische Eigenschaften den Erkennungsvektors ableiten:

Tabelle 18: Erkennungskriterien für PS00008

Position im Motiv	Zugelassene Energie
$\boldsymbol{R_1}$	$\mathrm{E}_1 < -20.0$
$\boldsymbol{R_2}$	$\mathrm{E}_2 < 1.0$
$\boldsymbol{R_3}$	$3.0 > \mathrm{E}_3 > -1.0$
$\boldsymbol{R_4}$	$\mathrm{E}_4 < -10.0$
$\boldsymbol{R_5}$	$\mathrm{E}_5 < -25.0$

Wird in einem Energieprofil ein Fragment gefunden, welches an allen Positionen, in den in Tabelle 19 aufgeführten, energetischen Grenzen liegt, so wird das Fragment als Kern betrachtet und um eine Position R_6 erweitert. Ist R_6 dann noch kleiner als -18.0 soll das erweiterte Fragment als PS00008 detektiert werden.

Mit diesen Kriterien wurde der Energieprofildatensatz (insgesamt 5567 Energieprofile) geparst. Der Testdetektor erkannte auf diesem Weg allerdings nur 685 Fragmente als PS00008-Motive. Definiert man das PS00008-Motiv lediglich als aus seinem Kern bestehend, kommt der Testdetektor auf eine Trefferzahl von 5304. Dies wären immerhin 36,3 % der 14612 tatsächlich vorhandenen PS00008-Motive. Es ist jedoch davon auszugehen, dass ein Detektor, der auf so einfachem Wege arbeitet eine sehr hohe Fehlerquote besitzt.

Die Frage, die sich an dieser Stelle auftut, ist die nach der grundsätzlichen Detektierbarkeit der Motive.

7 Ausblick

Im Folgenden soll ein Ausblick dahingehend gegeben werden, wie die weitere Arbeit mit strukturellen und funktionellen Motiven und deren Energievektoren aussehen könnte. Weiterhin soll ein kurzer Ausblick auf die mögliche energieprofilbasierte Arbeit mit Faltungsklassen gegeben werden.

7.1 Weitere Arbeit mit Motiven

Was nun noch im Kontext der Motive notwendig ist, ist es einen intelligenten Detektor zu entwickeln, der in der Lage ist die Motive mit einer hinreichend hohen Effizienz allein aus Energieprofildaten zu detektieren. Die Ergebnisse haben gezeigt, dass es ein geeigneter Ansatz wäre es nicht nur auf charakteristische Energievektoren hin zu überprüfen, sondern auch sequentielle sowie sekundärstrukturelle Eigenschaften zu berücksichtigen.

Ein weiterer Punkt, der in der Arbeit mit funktionellen Motiven interessant ist, ist die Frage nach der Ausübung ihrer Funktion. Wie binden beispielsweise Liganden an den Motiven? Wie verhält sich das Motiv räumlich und energetisch, während es aktiv in einen katalytischen Prozess eingebunden ist. Ein Verständnis dieser Thematik könnte einen noch wesentlicheren Einblick darauf gewähren, wie sich Evolution auf energetischer Ebene vollzieht.

7.2 Ausblick auf Faltungsklassen

Datenbanken wie SCOP und CATH bieten die Möglichkeit Proteinstrukturen eine Faltungsklasse zuzuordnen und so strukturell zu klassifizieren. Die Proteine werden auf der CATH beispielsweise nach einem hierarchischen System in Klassen, Architekturen, Topologien und Homologe Superfamilien eingeteilt:

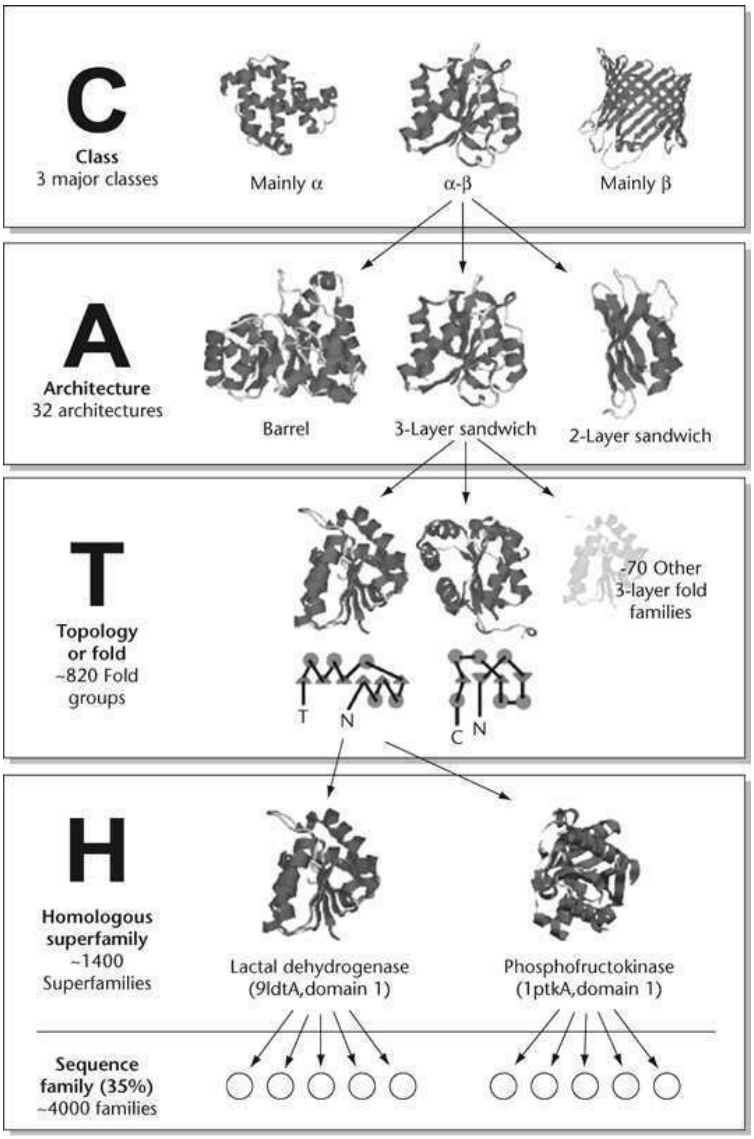


Abbildung 39: Einteilung von Strukturen nach CATH

Die CATH-Datenbank bietet ein Modell dafür an, wie Proteinstrukturen zu großen Strukturgruppen zusammengefasst werden können.

Diese Klassifizierung sollte sich auch in den Energieprofilen wiederspiegeln. Ein Beispiel hierfür soll die Struktur von 3KVN liefern.

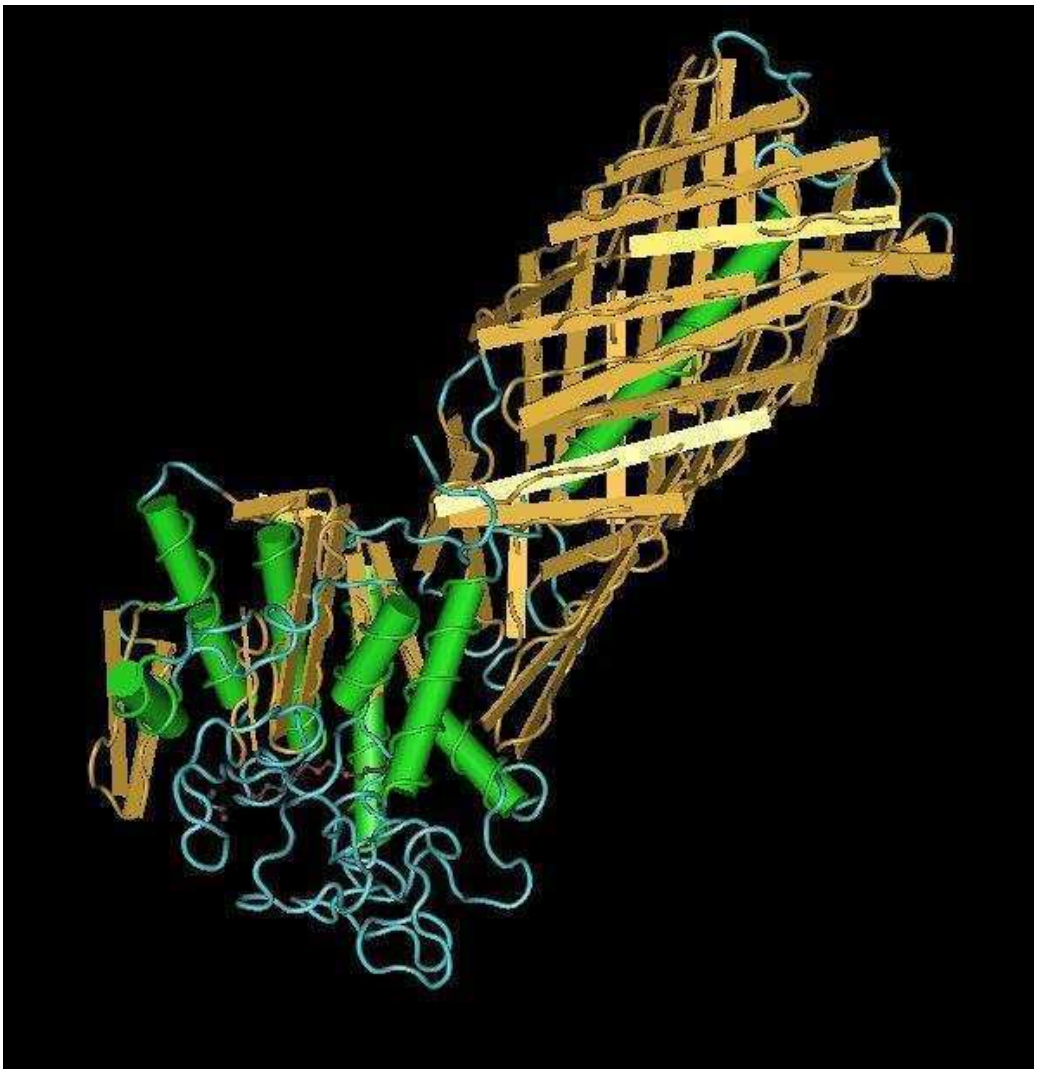


Abbildung 40: Struktur von 3KVN

3KVN ist nach der CATH-Klassifizierung ein Protein mit Betabarrel, in dem eine lange Helix lokalisiert ist.

Berechnet man das Energieprofil von 3KVN, so lässt sich vergleichsweise schnell der Bereich lokalisieren, in dem das Betabarrel über einige kurze Coil Regionen zusammen mit der zentralen Helix aufgebaut ist:

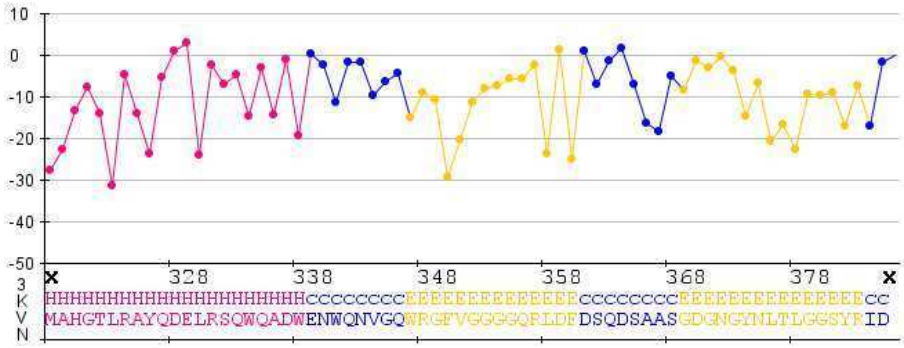


Abbildung 41: Energieprofilausschnitt von 3KVN

Ausgeschnitten aus dem Gesamtprofil ist hier der Energiebereich, in dem die zentrale Helix (rosa) über einen Coil-Bereich (blau) in die Faltblätter des Betabarrels übergeht.

Es wird schnell offensichtlich, dass zwar die Helix einen charakteristischen alternierenden Energieverlauf besitzt, die Coil- und Faltblattbereiche allerdings schwieriger voneinander zu differenzieren sind. Die Detektion dieser Strukturelemente, die für die Einordnung von 3KVN als Betabarrel vonnöten wäre würde an den Übergängen der Strukturelemente problematisch werden.

Die Grenzen der Sekundärstrukturelemente zueinander sind im Energieverlauf schwierig zu setzen. Auch wenn beachtet wird, dass coil-Bereiche meist durch eine hohe Energie realisiert sind, lässt sich beispielsweise an den Enden von Helices nur schwer darauf schließen, wo die Helix aufhört und der Coil beginnt.

8 Zusammenfassung

Die Analysen der strukturellen und funktionellen Proteinmotive haben offenbart, dass sich Motive, obgleich sie strukturell beziehungsweise funktionell klar definiert sind, auf Ebene der Energieprofile weiter differenzieren lassen. Dieser Umstand hat Ursachen in der sequentiellen Realisierung eines Strukturmotivs beziehungsweise in der strukturellen Realisierung eines funktionellen Motivs.

Es ist dabei festgestellt worden, dass Motive, die beispielsweise strukturell klar voneinander abzugrenzen sind, sich auf energetischer Ebene in ihren Energieverläufen gleichen können. Weiterhin kann gesagt werden, dass strukturelle Motive in der Lage sind funktionelle Motive aufzubauen oder zumindest im Hinblick auf die Ebene der Sekundärstruktur sinnvoll mit diesen assoziiert sind.

Obgleich viele Motive charakteristische Energieverläufe besitzen ist eine Implementierung eines Motivdetektors noch ausstehend. Die Gründe hierfür werden an vielen Stellen der vorliegenden Arbeit ersichtlich. Die Genauigkeit eines Algorithmus, der in der Lage ist Motive allein aus Energieprofildaten zu detektieren liegt nach der Meinung des Autors in einem Bereich von 50-70%.

Schlussendlich wurde ein Ausblick dahingehend gegeben, wie die weitere Arbeit mit Struktur- und Funktionsmotiven und das Herangehen an die Thematik der energieprofilbasierten Analyse von Faltungsklassen aussehen könnte.

Literatur

[1] A. Marisco, A. Tuukkanen, D. Labudde, F. Dressel, M. Schroeder, *Understanding of SMFS barriers by means of protein energy profiles*, Proc. GCB.

[2] A. Golovin, K. Henrick, *Exploring Protein Sites and Motifs, BMC Bioinformatics*, 9:312, 2008.

[3] S. Jones, D. Jones, A. Mitchie, C. Orengo, M. Swindells, J. Thornton, *CATH - a hierarchic classification of protein domain structures, Structure 5(8), 1093-1108, 1997.*

[4] B. Alberts, D. Bray, A. Johnson, J. Lewis, M.Raff, K. Roberts, P. Walter, Lehrbuch der Molekularen Zellbiologie, 2. Auflage, Weinheim, WILEY-VCH, 2001.

[5] URL: <http://daten.didaktikchemie.uni- bayreuth.de/umat/proteine /aminosaeure.gif>

[6] URL: <http://www.dsimb.inserm.fr/~debrevern/VENN_DIAGRAM/ aa_venn_diagram.png>

[7] URL: <http://daten.didaktikchemie.uni- bayreuth.de/umat/peptidsynthese/peptidbindung.gif>

[8] P. Y. Bruice, *Organic Chemistry*, Pearson Education Inc., 4. Auflage, S. 960–962, 2004

[9] URL: <http://www.wiley-vch.de/HOME/bioinformatik/prot/Prot_1d/ueb/pepbind.png>

[10] F. Heinke, Energieprofilbasierende Analysemethoden von Proteinfamilien, Mittweida, 2010.

[11] R. Merkl, S. Waack, Bioinformatik Interaktiv, 1. Auflage, Weinheim, WILEY-VCH, 2003

[12] URL: <http://wps.prenhall.com/wps/media/objects/602/616516/Media_Assets/Chapter24/Text_Images/FG24_07.JPG>

[13] URL: < http://upload.wikimedia.org/wikipedia/commons/2/25/Bsheet.gif>

[14] G.N. Ramachandran,C. Ramakrishnan, V. Sasisekharan, *Stereochemistry of polypeptide chain configurations*, Journal of Molecular Biology 7: 95–9, 1963

[15] abgeändert nach URL: <http://www.chemgapedia.de/vsengine/media/vsc/de/ch/8/bc/proteine/aminos_u_einleit/gif_pdb_mov/ramachandran4_swf_altref.jpg>

[16] K. Kühn, Struktur und Biochemie des Kollagens, Chemie unserer Zeit, 8: 97–103, 1974

[17] W. Kabsch, C. Sander, *How good are predictions of protein secondary structure?,* FEBS Lett 155 (2): 179–82, 1983

[18] B. Kumar, N. Jani, *Prediction of Protein Secondary Structure based on GOR Algorithm Integrating with Multiple Sequences Alignment*, International Journal of Advanced Engineering & Applications, 2010

[19] D. T. Jones, *Protein secondarystructureprediction based on position-specific scoring matrices*, Journal of Molecular Biology, Vol. 292, Issue 2: 195–202, 1999

[20] URL: <http://www.guidobauersachs.de/oc/disulf.gif>

[21] URL: <http://www.nibsc.uk/science/protein_hormones__endocrine_p.aspx>

[22] URL: <http://www.lsbu.ac.uk/water/protein2.html>

[23] H. Wegele, L. Muller, J. Buchner: *Hsp70 and Hsp90 - a relay team for protein folding*, Rev. Physiol. Biochem. Pharmacol, Springer Verlag, 2004

[24] A.J. Shepherd, D. Gorse, J.M. Thornton, *Prediction of location and type of beta-turns in proteins using neural network***,** Protein Science 8(5):1045-55, 1999

[25] D. J . Craik, N. L. Daly, C. Waine, *The cystine knot motif in toxins and implications for drug design*, Toxicon, 39(1):43-60, 2001

[26] URL: <http://weblogo.berkeley.edu/>

[27] E. Milner-White, J. Russell, *Sites for Phosphates and Iron-Sulfur Thiolates in the First Membranes: 3 to 6 Residue Anion-Binding Motifs (Nests),* Orig Life Evol Biosph, 35(1):19-27, 2005

[28] URL: <http://prosite.expasy.org/prosuser.html#meth1>

[29] D. Baker, O. Grana, R.M. MacCallum, J. Meiler, M. Punta, B. Rost, A. Valencia, *CASP6 assessment of contact prediction*, Proteins, 61 Suppl 7:214-224, 2005

[30] Adaptiert nach F. Heinke, 2012

[31] J.G. Henikoff, S. Henikoff, *Amino acid substitution matrices from protein blocks,* Biochemistry, Vol. 89, pp. 10915-10919, 1992

[32] URL: <http://upload.wikimedia.org/wikipedia/en/3/37/Tree_UPGMA.png>

[33] M. Thomas, K. Schulten, *A neural gas network learns topologies*, Artificial Neural Networks. Elsevier. pp. 397-402, 1991

[34] F. Curatelli, O. Mayora-Iberra, *Competitive learning methods for efficient Vector Quantizations in a speech recognition environment*, Advances in artificial intelligence, Mexican International Conference on Artificial Intelligence, Acapulco, 2000

[35] A. Angelopoulou, A. Psarrou, J.G.Rodriguez, K. Revett, *Automatic landmarking of 2D medical shapes using the growing neural gas network*, Computer vision for biomedical image applications: First international workshop, Beijing, 2005

[36] URL: < http://www.ebi.ac.uk/pdbe-site/pdbemotif/>

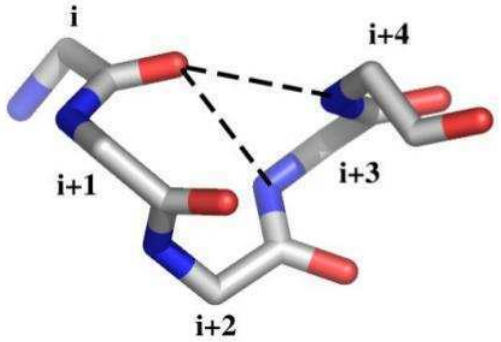
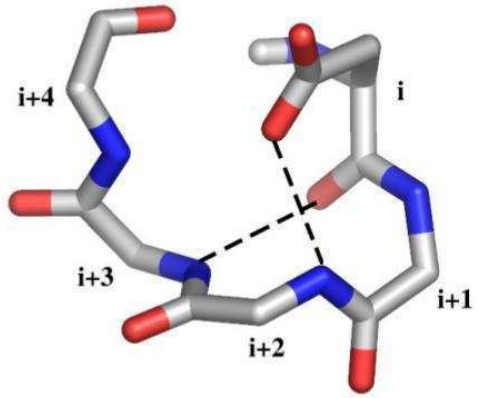
[37] URL: <http://bioservices.hs-mittweida.de/Epros/Index?page=dataset >

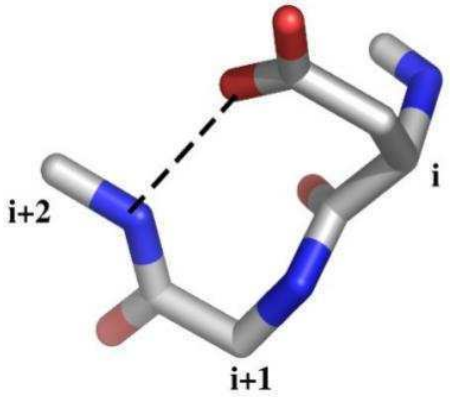
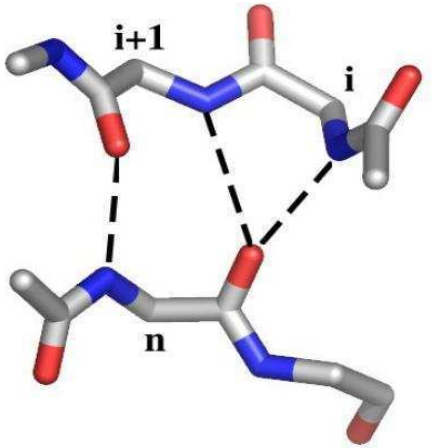
[38] URL: <http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html>

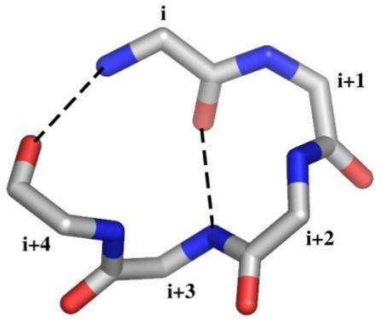
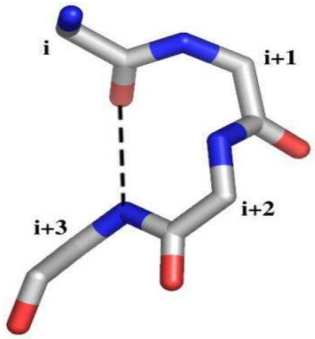
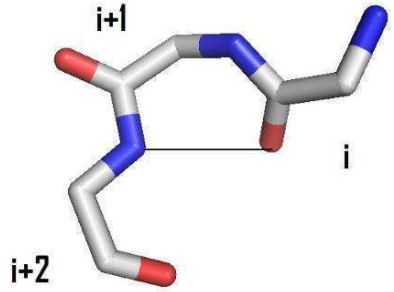
[39] URL: <ttp://www.cbrc.jp/pdbreprdb-cgi/reprdb_menu.pl>

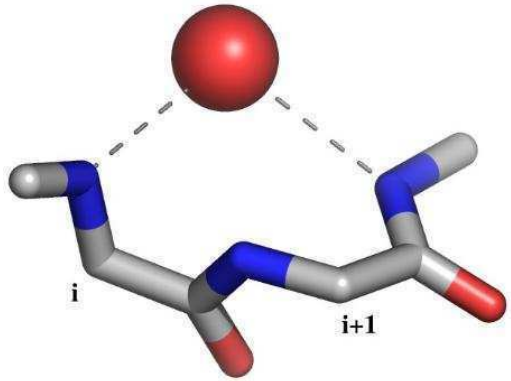
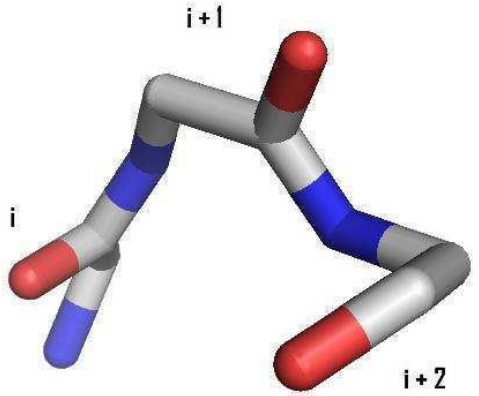
Anlagen

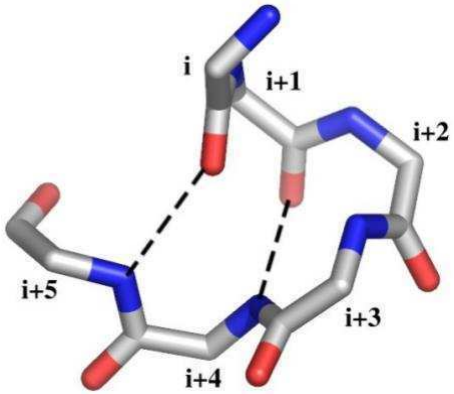
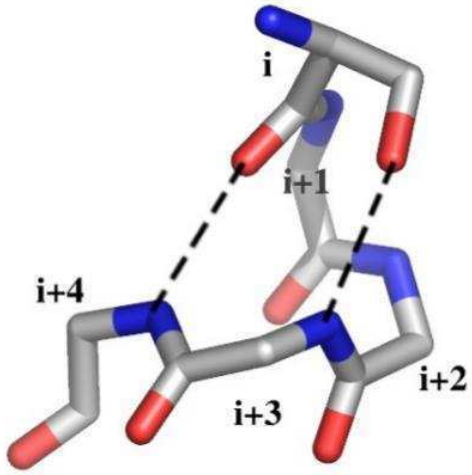
Anlagen, Teil 1 – Erläuterungen zu Strukturmotiven

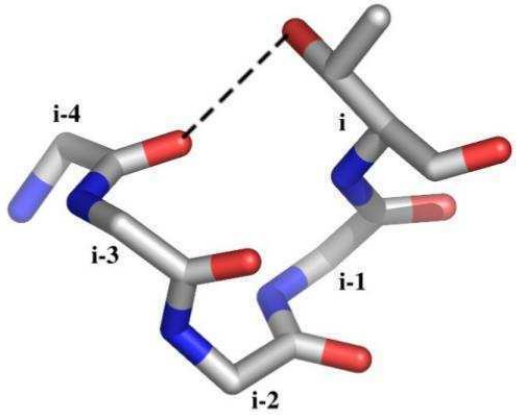
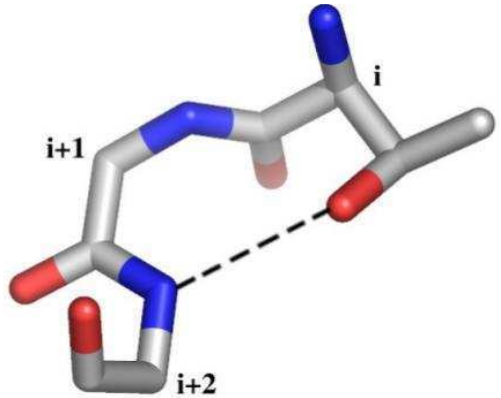
Motivbezeichnung	Länge in AS	Anzahl an Subtypen	Anzahl an H-Bonds	Lokalisation der H-Bonds	Besonderheit	Strukturelle Darstellung
Alpha-Beta motif	5	0	2	1.: $R_i – R_{i+3}$ 2.:$R_i – R_{i+4}$	Die φ-Winkel der AS R_{i+1}, R_{i+2} und R_{i+3} sind negativ, deshalb handelt es sich um ein von Natur aus nach links orientiertes Motiv.	
Asx-motif	5	0	2	1.: $R_i – R_{i+2}$ v $R_i – R_{i+3}$ 2.: $R_{i+3} – R_{i+4}$	Die erste Position R_i dieses Motivs ist stets durch die Aminosäuren Aparagin oder Asparaginsäure realisiert.	

Motivbezeichnung	Länge in AS	Anzahl an Subtypen	Anzahl an H-Bonds	Lokalisation der H-Bonds	Besonderheit	Strukturelle Darstellung
Asx-turn	3	4	1	$R_i – R_{i+2}$	Auch hier ist die erste Position durch die Aminosäuren Asparagin bzw. Asparaginsäure determiniert. An den Enden von Asx-turns sind oftmals Calcium-Ionen koordiniert.	
Beta-bulge	2	0	3	1.: $R_i – n$ 2.: $R_{i+1} – n$ 3.: $R_{i+1} – n$	Bestandteil eines β-Sheets, in dem zwei Aminosäuren über drei H-Bonds mit einem dritten Residuum n einer anderen Sequenzregion verbunden sind.	

Motivbezeich-nung	Länge in AS	Anzahl an Subtypen	Anzahl an H-Bonds	Lokalisation der H-Bonds	Besonderheit	Strukturelle Darstellung
Beta-bulge loop	5 oder 6	2	2	Typ I: 1.: $R_i – R_{i+3}$ 2.: $R_i – R_{i+4}$ Typ II: 1.: $R_i – R_{i+4}$ 2.: $R_i – R_{i+5}$	Ebenfalls ein in β-Sheets enthaltenes Motiv.	
Beta-turn	4	4	0 oder 1	$R_i – R_{i+3}$		
Gamma-turn	3	2	1	$R_i – R_{i+2}$	Kann in Gamma-classic Type und Gamma-turn inverse Type unterschie-den werden.	

Motivbezeichnung	Länge in AS	Anzahl an Subtypen	Anzahl an H-Bonds	Lokalisation der H-Bonds	Besonderheit	Strukturelle Darstellung
Nest	3	2	0		Ein Nest enthält an keiner Stelle die Aminosäure Prolin. Nests nehmen eine Schlüsselposition in Phosphat-bindenden Stellen ein, um beispielsweise ATP oder GTP zu binden.	
Niche	3 oder 4	4	0		Die erste Position eines Niche-Motivs kann jede mögliche Winkelkonformation einnehmen. In der Literatur wird dieses Motiv gelegentlich als Catmat bezeichnet.	

Motivbezeich-nung	Länge in AS	Anzahl an Subtypen	Anzahl an H-Bonds	Lokalisation der H-Bonds	Besonderheit	Strukturelle Darstellung
Schellmannloop	6	0	2	1.: $R_i – R_{i+4}$ 2.: $R_i – R_{i+5}$	Beschrieben sind auch Schellmannloops der Länge 7, diese wurden jedoch aufgrund fehlender Daten aus der Betrachtung herausgenommen	
St-motif	5	0	2	1.: $R_i – R_{i+2}$ v Ri+3 2.: $R_i – R_{i+3}$ v R_{i+4}	Die erste Position R_i dieses Motivs ist stets durch die Aminosäuren Serin oder Threonin realisiert.	
Motivbezeich-	**Länge**	**Anzahl an**	**Anzahl an**	**Lokalisation**	**Besonderheit**	**Strukturelle Darstellung**

nung	in AS	Subtypen	H-Bonds	der H-Bonds		
St-staple	5	0	1	$R_i – R_{i+3}$ v R_{i+4}	Die φ-Winkel der AS R_{i-1}, R_{i-2} und R_{i-4} sind negativ, deshalb handelt es sich auch hier um ein von Natur aus links orientiertes Motiv. Das erste Residuum ist ebenfalls durch Serin oder Threonin realisiert.	
St-turn	3	4	1	$R_i – R_{i+2}$	Wie bei allen St-Motiven ist die erste Position im Motiv durch Serin oder Threonin determiniert.	

Anlagen, Teil 2 – Erläuterungen zu funktionellen Motiven

Motivbe-zeichnung	Beschreibung	Länge in AS	Sequenzpat-tern	Funktion
PS00001	N-Glykolisationsstelle	4	N-{P}-[ST]-{P}	Bindet ein Zuckermolekül an Proteinoberfläche
PS00004	cAMP-Phosphorylisationsstelle	4	[RK] (2)-x-[ST]	cAMP- oder cGMP-abhängige Protein-Kinase-Phosphorylisationsstelle
PS00005	Phosphorylisationsstelle	3	[ST]-x-[RK]	Protein Kinase C Phosphory-lisationsstelle
PS00006	Phosphorylisationsstelle	4	[ST]-x(2)-[DE]	Casein Kinase II Phosphory-lisationsstelle
PS00007	Phosphorylisationsstelle	7 - 9	[RK]-x(2,3)-[DE]-x(2,3)-Y	Tyrosin Kinase Phosphoryli-sationsstelle
PS00008	N-Myristoylierungsstelle	6	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}	Bindet Myristinsäure an den N-terminalen Gylcinrest von Proteinen
PS00009	Amidationsstelle	4	x-G-[RK]-[RK]	Bindet Amide an das C-terminale Ende von Protei-nen
PS00016	Zelladhäsionsstelle	3	R-G-D	Stelle zur Zelladhäsion
PS00017	ATP-/GTP-Bindungsstelle	8	[AG]-x(4)-G-K-[ST]	ATP-/GTp-bindende Stelle in einem P-loop
PS00107	ATP-Bindestelle	9 - 22	x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K	Protein Kinase ATP-Bindungsstelle
PS00108	Protein Kinase aktive Stelle	13	[LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3)	Serin/Threonin Protein Kina-se aktive Stelle

Anlagen, Teil 3 – Verteilung der Sekundärstruktur-elemente in Motiven

Motiv	c/H/S/G 1. AA	c/H/S/G 2. AA	c/H/S/G 3. AA	c/H/S/G 4. AA	c/H/S/G 5. AA	c/H/S/G 6. AA
Alphabeta-motif	0,35/1,54/0,23/2,12	0,17/1,69/0,02/1,88	0,21/1,66/0,00/1,87	0,29/1,60/0,02/1,92	0,42/1,49/0,23/2,14	
Asx-motif	1,06/0,91/1,37/3,34	0,81/1,24/0,11/2,15	0,75/1,28/0,06/2,10	0,67/1,31/0,28/2,26	0,58/1,32/0,78/2,67	
Asx-turn	2,09/0,15/1,80/4,04	2,10/0,32/0,27/2,70	1,98/0,36/0,63/2,98			
Betabulge	0,44/0,00/12,90/13,34	0,08/0,00/15,04/15,12				
Betabule-loop (5)	1,98/0,00/3,73/5,71	2,59/0,00/0,13/2,72	2,61/0,00/0,00/2,61	2,59/0,00/0,10/2,69	2,18/0,00/2,55/4,73	
Betabulge-loop (6)	0,57/0,01/12,07/12,65	2,23/0,24/0,17/2,64	2,25/0,25/0,00/2,50	2,24/0,25/0,06/2,55	2,45/0,10/0,11/2,66	0,76/0,00/11,00/11,76
Beta-turn	0,89/0,58/5,24/6,71	1,60/0,68/0,13/2,42	1,68/0,64/0,05/2,37	1,03/0,54/4,71/6,29		
Gamma-turn	2,38/0,03/1,07/3,49	2,61/0,00/0,00/2,61	2,52/0,06/0,00/2,58			
Nest	2,23/0,26/0,04/2,53	2,19/0,28/0,04/2,52	1,93/0,35/1,02/3,31			
Niche (3)	2,17/0,09/1,81/4,08	2,42/0,08/0,45/2,95	2,13/0,19/1,25/3,57			
Niche (4)	1,96/0,21/2,07/4,24	2,19/0,13/1,35/3,68	2,41/0,09/0,46/2,96	2,17/0,03/2,32/4,53		
Schellmann-loop	0,26/1,56/0,66/2,47	0,27/1,57/0,43/2,27	0,45/1,45/0,40/2,30	1,28/0,89/0,32/2,48	1,57/0,63/0,77/2,98	1,36/0,64/1,89/3,90
St-motif	1,26/0,77/1,43/3,46	0,46/1,49/0,00/1,95	0,46/1,49/0,00/1,95	0,42/1,50/0,16/2,08	0,39/1,50/0,32/2,21	
St-staple	0,19/1,67/0,07/1,93	0,12/1,72/0,01/1,86	0,10/1,74/0,01/1,85	0,12/1,72/0,03/1,87	0,27/1,61/0,10/1,98	
St-turn	1,71/0,39/1,97/4,08	2,06/0,36/0,21/2,62				

	c	H	S
Gesamtverteilung [%]	38,33	55,24	6,44

Motiv	c/H/S/G 1. AA	c/H/S/G 2. AA	c/H/S/G 3. AA	c/H/S/G 4. AA	c/H/S/G 5. AA	c/H/S/G 6. AA
PS00001	1,17/0,83/1,06/3,06	1,16/0,84/1,06/3,06	1,06/0,88/1,13/3,08	1,00/0,91/1,18/3,09		
PS00004	0,98/1,00/1,04/3,01	0,97/0,98/1,08/3,03	1,02/0,96/1,04/3,02	0,96/1,00/1,07/3,02		
PS00005	0,96/1,01/1,04/3,01	0,96/1,02/1,02/3,00	0,96/1,02/1,02/3,00			
PS00006	0,94/1,10/0,90/2,94	0,93/1,14/0,84/2,91	0,92/1,16/0,82/2,90	0,91/1,16/0,83/2,90		
PS00007	0,89/1,12/0,95/2,95	0,90/1,12/0,92/2,94	0,94/1,10/0,89/2,94	0,94/1,09/0,92/2,95	0,89/1,11/0,97/2,96	0,86/1,10/1,02/2,98
	7. AA	8. AA	9. AA			
	0,87/1,05/1,11/3,03	0,86/1,04/1,14/3,04	0,82/1,13/1,04/2,98			
	1. AA	2. AA	3. AA	4. AA	5. AA	6. AA
PS00008	1,18/0,88/0,97/3,02	1,13/0,87/1,04/3,04	1,02/0,92/1,12/3,06	0,97/0,95/1,14/3,06	0,93/0,99/1,12/3,04	0,91/1,01/1,11/3,04
PS00009	1,18/0,90/0,92/3,00	1,28/0,85/0,87/2,99	1,21/0,83/0,99/3,04	1,01/0,83/1,32/3,15		
PS00016	1,02/0,89/1,18/3,09	1,10/0,93/0,98/3,01	1,17/0,92/0,90/2,98			
PS00017	1,24/0,67/1,27/3,17	1,53/0,69/0,77/2,99	1,52/0,72/0,75/2,98	1,52/0,70/0,77/2,99	1,46/0,76/0,75/2,97	0,86/1,26/0,72/2,84
	7. AA	8. AA				
	0,70/1,34/0,81/2,85	0,57/1,41/0,88/2,86				
	1. AA	2. AA	3. AA	4. AA	5. AA	6. AA
PS00107	0,96/0,00/2,98/3,95	0,72/0,00/3,36/4,08	0,72/0,00/3,36/4,08	1,44/0,19/1,87/3,50	1,68/0,19/1,49/3,37	1,93/0,19/1,12/3,24
	7. AA	8. AA	9. AA	10. AA	11. AA	12. AA
	0,96/0,19/2,61/3,77	0,48/0,19/3,36/4,03	0,96/0,00/2,98/3,95	0,96/0,00/2,98/3,95	0,72/0,00/3,36/4,08	0,48/0,00/3,73/4,21
	13. AA	14. AA	15. AA	16. AA	17. AA	18. AA
	0,24/0,00/4,10/4,34	0,48/0,00/3,73/4,21	1,93/0,00/1,49/3,42	1,93/0,00/1,49/3,42	1,68/0,00/1,87/3,55	1,68/0,00/1,87/3,55
	19. AA	20. AA	21. AA	22. AA	23. AA	24. AA
	1,44/0,19/1,87/3,50	0,24/0,19/3,73/4,17	0,24/0,39/3,36/3,99	0,48/0,39/2,98/3,85	0,53/0,42/2,85/3,80	0,26/0,42/3,26/3,94
	1. AA	2. AA	3. AA	4. AA	5. AA	6. AA
PS00108	0,21/1,00/2,24/3,44	0,83/0,83/1,60/3,25	1,86/0,33/0,96/3,15	1,65/0,33/1,28/3,26	1,86/0,17/1,28/3,30	1,44/0,17/1,92/3,53
	7. AA	8. AA	9. AA	10. AA	11. AA	12. AA
	1,03/0,83/1,28/3,14	1,44/1,00/0,32/2,76	1,24/1,00/0,64/2,87	1,86/0,33/0,96/3,15	1,24/0,33/1,92/3,49	0,62/0,50/2,56/3,68
	13. AA					
	0,62/0,50/2,56/3,68					

	c	H	S
Gesamtverteilung [%]	34,63	43,03	22,34

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, den 22.8.201

Oswald, Silvio